A Tree-based Approach to Integrated Action Localization, Recognition and Segmentation

Third Workshop on Human Motion Understanding, Modeling, Capture and Animation

Zhuolin Jiang[1], Zhe Lin[2], Larry S. Davis[1]

[1]University of Maryland at College Park
[2]Adobe Systems Incorporated

Sep. 10th, 2010

UMIACS
University of Maryland Institute for Advanced Computer Studies

Goal & Challenges

- Goal
  - Propose a simultaneous approach to localize and recognize multiple action classes based on a unified tree-based framework under moving camera and dynamic backgrounds

- Challenges
  - Dynamic backgrounds
    (e.g. moving people, vehicles)
  - Moving camera
  - Occlusions
  - Appearance and illumination variations

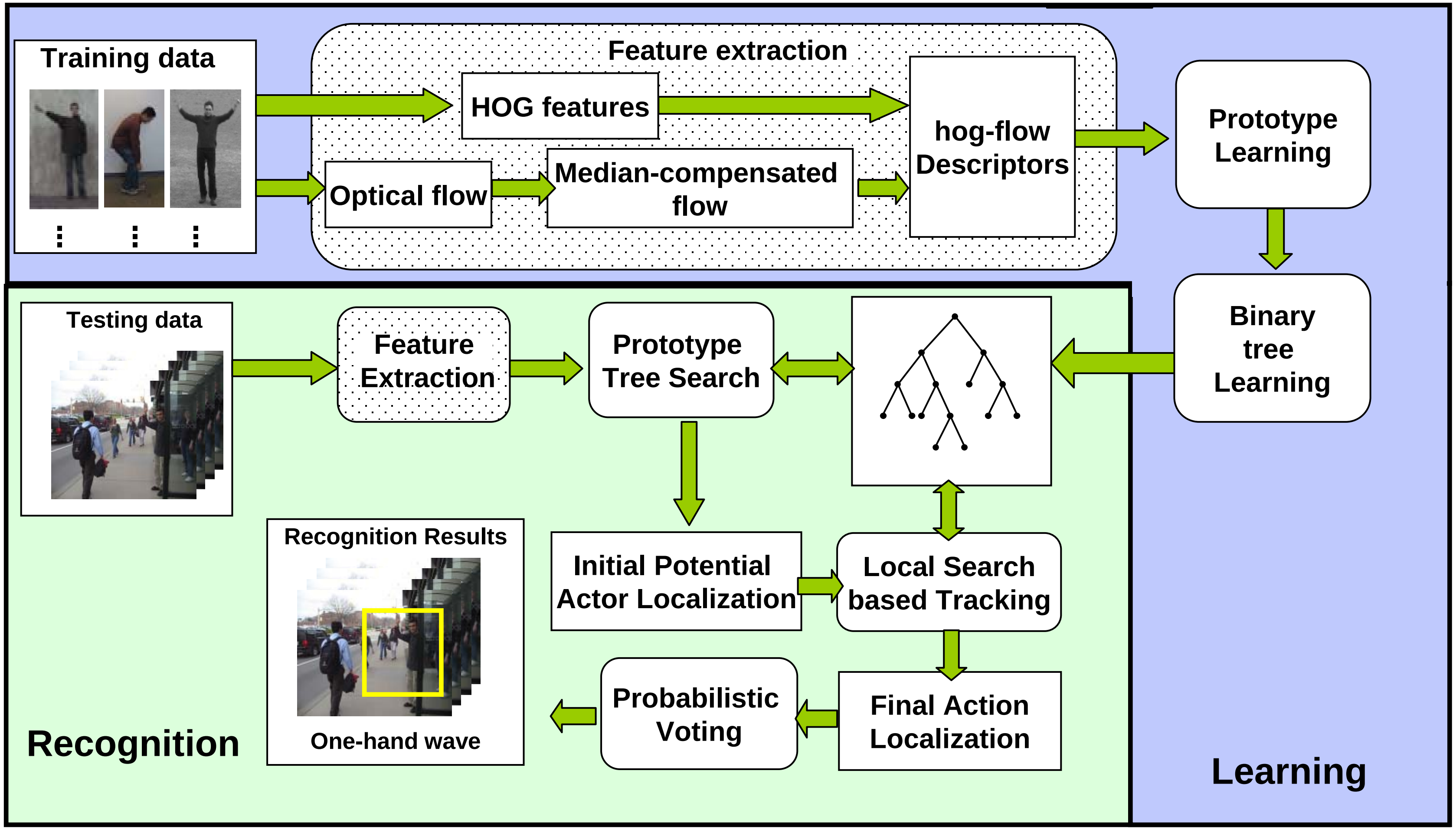
Extension of our previous approach

- HOG-based shape feature
- The prototype tree is used to simultaneously localize and recognize actions
- The probabilistic framework is constructed to determine action category labels and action prototypes

Contributions

- A simplified HOG-based shape feature is adopted to enhance the joint shape-motion descriptors proposed in our previous approach.
- A binary prototype tree based approach is introduced to efficiently localize and recognize multiple action classes.
- Action Recognition is model as a maximum probability estimation problem of the conditional probabilities of action category labels and action prototypes.

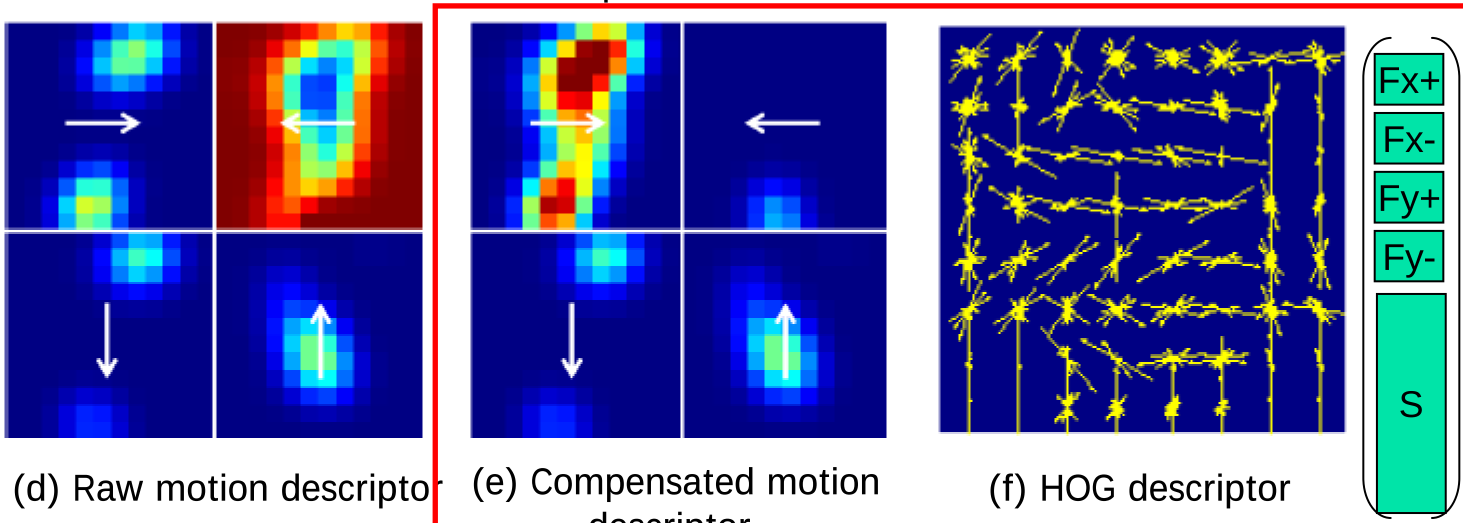
Overview



Training data
Feature extraction
HOG features
Optical flow
Median-compensated flow
hog-flow Descriptors
Prototype Learning
Binary tree Learning
Testing data
Feature Extraction
Prototype Tree Search
Initial Potential Actor Localization
Local Search based Tracking
Final Action Localization
Probabilistic Voting
Recognition Results
One-hand wave
Recognition
Learning

Action Representation by Hog-flow Descriptors

(a) Raw optical flow (b) Motion compensated optical flow (c) Image spatial gradient



(d) Raw motion descriptor (e) Compensated motion descriptor (f) HOG descriptor

Tree Model Construction and Matching

- Action Prototypes
  - k-means clustering
- Tree Model Construction
  - Hierarchical k-means clustering
  - Parameter learning
    - All tree nodes: Rejection thresholds $\Theta = (\theta_1, \theta_2, ..., \theta_{n_t})$
    - Leaf node λ_i includes:
      - A rejection threshold θ_i
      - A class distribution vector $\Omega_i = (\omega_{i,1}, ..., \omega_{i,m})$
      - Training frame indices matching with λ_i

$\theta_i = \tau D_{leaf_i}$

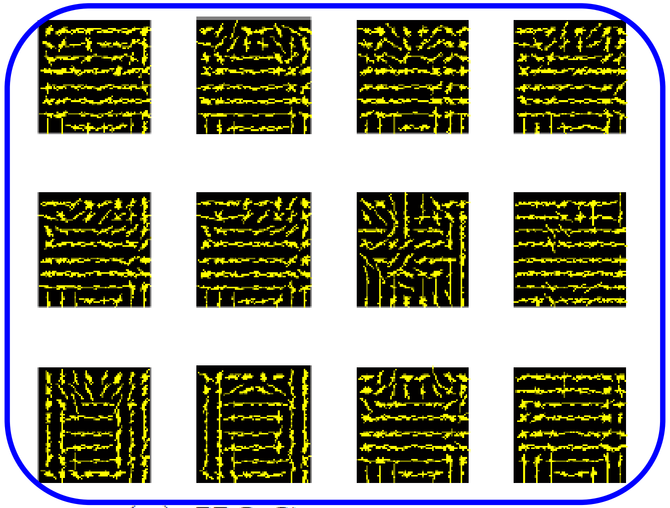
D_{leaf_i} - maximum Euclidean distance between tree node and children leaf nodes

$\hat{\Omega}_i = (\hat{\omega}_{i,1}, ..., \hat{\omega}_{i,m})$ $\hat{\omega}_{i,m} = \frac{F_{i,m}}{F_m}$

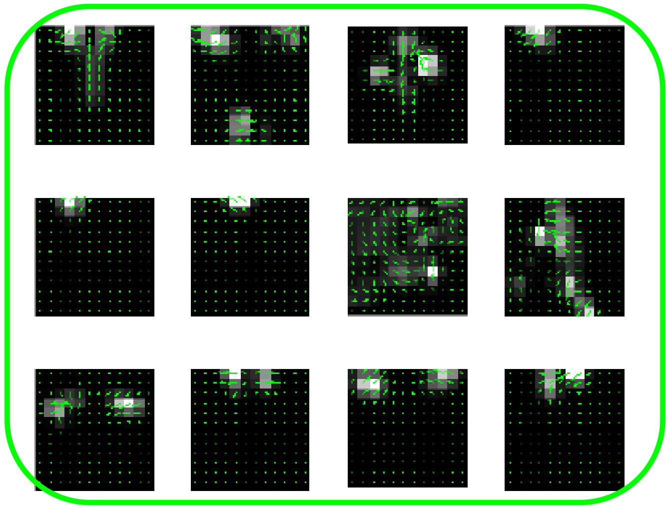
$F_{i,m}$ - number of training features from class m matching to λ_i

F_m - number of training features in class m

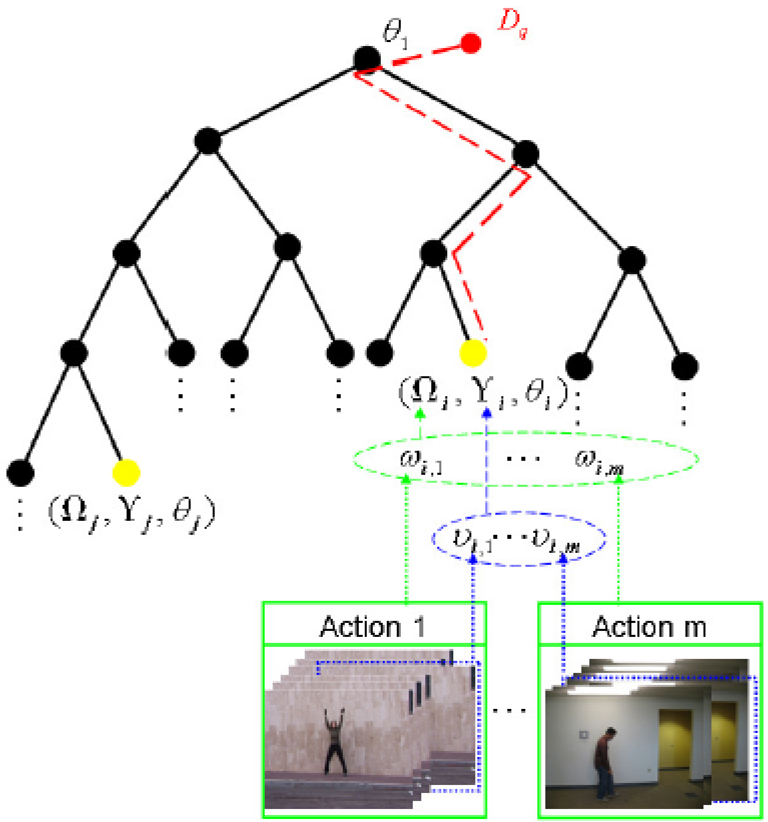
Action Prototype Tree



(a) HOG components



(b) Flow components



(c) Learned binary tree model

Action Recognition

- Conditional probability model

$$p(\alpha, \lambda, \beta|V) \propto \underbrace{p(\alpha|V, \lambda)}_{\text{action category mapping term}}\underbrace{p(\lambda|V, \beta)}_{\text{Prototype matching term}}$$

$\omega_{i,\alpha}$ in class distribution vector Ω_i

$$p(\lambda|V, \beta) \propto e^{-d(D(V,\beta),D(\lambda))}$$

V - observation r. v.
λ - prototype r. v. chosen from $\Lambda = (\lambda_1, \lambda_2 ... \lambda_k)$
β - actor location r. v. comprising image locaion (x,y) and scale s.
α - action category r.v. chosen from $A = (\alpha_1, \alpha_2 ... \alpha_m)$

- Optimization problem

$$(\alpha^*, \lambda^*, \beta^*) = \arg\max_{\alpha,\lambda,\beta} \boldsymbol{p}(\alpha, \lambda, \beta \,|\, \boldsymbol{V})$$

Action Recognition

- Conditional probability optimization
  - Given actor location β_t and observation *V* at frame t, we define a score function for the corresponding prototype $\lambda_i(\beta_t)$ and action category label α_i at frame t as:

$$J_t(\alpha_i) = \omega_{i,\alpha_i} e^{-d(D(V,\beta_t),D(\lambda_i(\beta_t)))},$$

  - The optimal action label is:

$$\alpha_i^* = argmax_{\{\alpha_i\}_{i=1\dots m}} \sum_{t=l_{start}}^{l_{end}} J_t(\alpha_i),$$

Action Segmentation

- Segmentation mask
  - After tree construction, each action category in *i*-th leaf node (prototype) has its own set of representative binary silhouettes $\{b_j\}_{j=1..m}$, which is identified by $\Upsilon_i = (\upsilon_{i,1}, ..., \upsilon_{i,m})$. The segmentation mask for *i*-th leaf node is defined as:

    $B_i = \sum_{j=1}^{m} \omega_{i,j} b_j$

Experiments

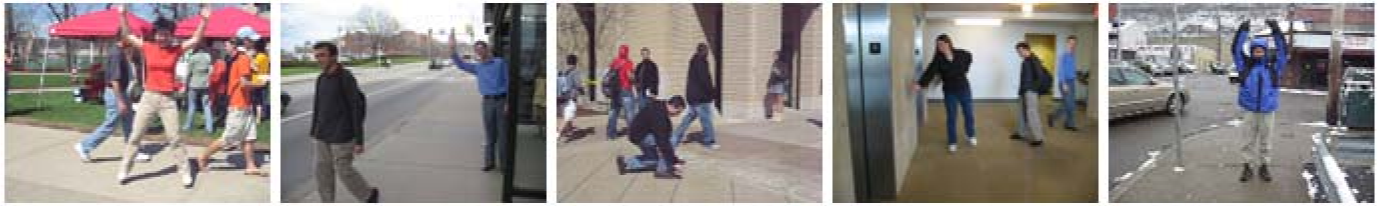
- Hog-flow Descriptor
  - 8*8*9 hog descriptor
  - Four channels of 12*12 motion descriptor
  - Total dimension: 1152

Datasets

- CMU Action Dataset

  http://www.yanke.org/research.htm

  - 48 training video sequences,110 testing video sequences
  - 5 action classes
  - Hand-held camera, dynamic backgrounds (moving persons or vehicle in the background)



- Weizmann Action Dataset

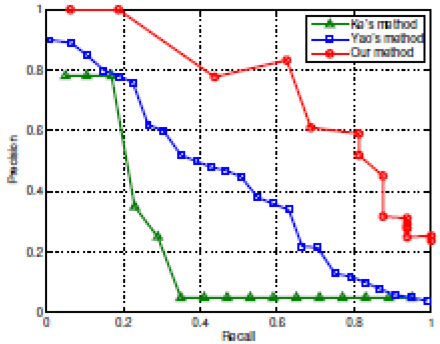
  http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

The Weizmann action dataset contains 90 videos of 10 actions performed by 9 individuals.

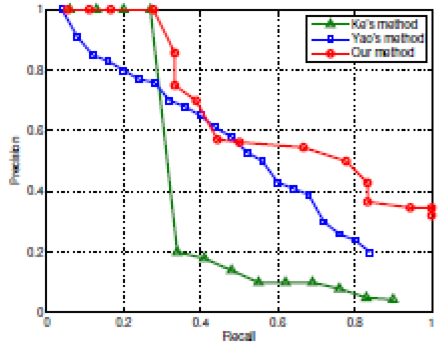
L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In IEEE Trans. PAMI, 29(12):2247–2253, 2007.

Results on the CMU Dataset

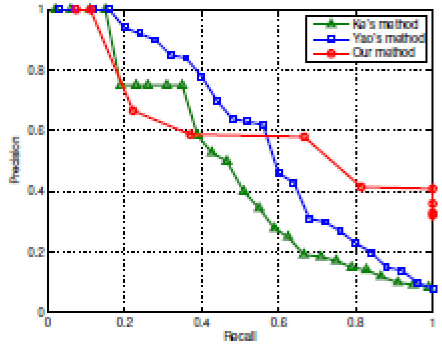
method	recog. rate (%)	avg. time (s)
500 proto.	84.55	0.86
1000 proto.	89.09	0.91
1700 proto.	89.09	0.88
2300 proto.	90	0.92
2789 proto.	100	0.89



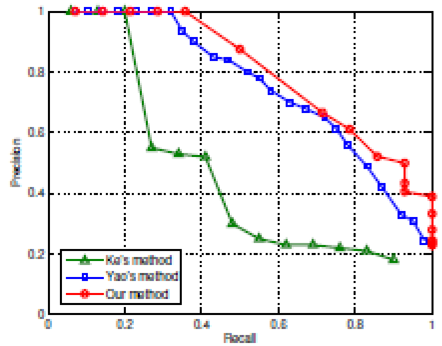
(a) jumping-jacks



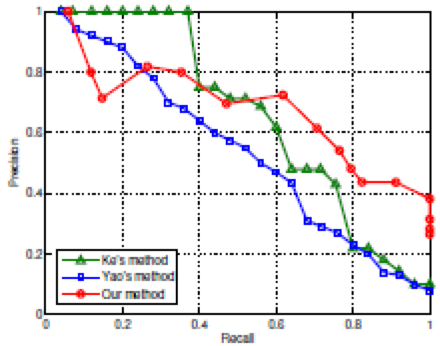
(b) one-handed-wave



(c) pick-up



(d) push-button



(e) two-handed-wave

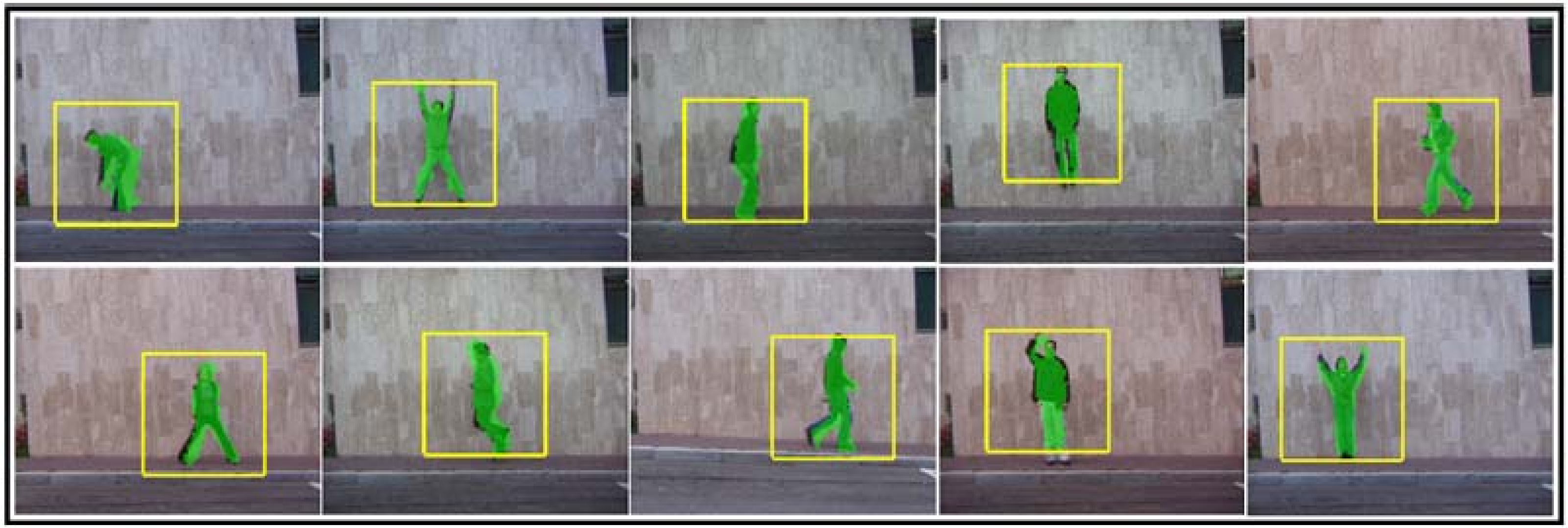
Results on the CMU Dataset

- Sliding Window-based Detection Candidates (number of Det.=20)

Results on the CMU Dataset

Results on the Weizmann Dataset

method	recog. rate (%)	avg. time (s)
500 proto.	91.11	0.91
1500 proto.	92.22	0.93
2500 proto.	88.89	0.94
3000 proto.	90.00	0.96
4000 proto.	94.44	0.96
all descriptors	100	0.94
Fathi [16]	100	N/A
Schindler [23]	100	N/A
Lin [9]	100	N/A
Jhuang [22]	98.8	N/A
Blank [1]	99.61	N/A
Thurau [25]	94.40	N/A



Demo on the CMU dataset

Summary

- Conclusions
  - The approach can yield good results for action localization and recognition in realistic scenarios with cluttered, dynamic backgrounds
  - The approach does not rely on background subtraction
- Future work
  - Incorporation of scene-specific cues or high-level spatial or temporal contexts would make the approach more reliable and accurate.
  - Extending the approach to Handle more challenging cases such as the presence of multiple interested actions performed simultaneously by multiple actors

Thank you!