Exploring Motion Boundary based Sampling and Spatial-Temporal Context Descriptors for Action Recognition

Xiaojiang Peng[1,2], Yu Qiao[2,3], Qiang Peng[1] and Xianbiao Qi[2]

[1]Southwest Jiaotong University, Chengdu, China
[2]Shenzhen Institutes of Advanced Technology, CAS, China
[3]The Chinese University of Hong Kong, China

Introduction

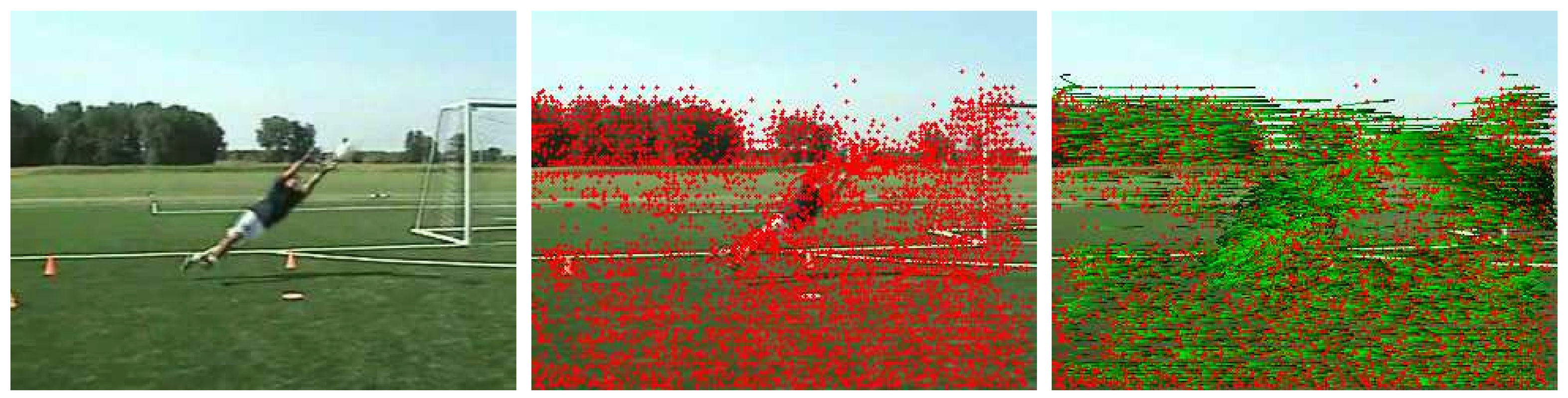
- **Goal**: Design new sampling strategy and descriptors to improve the recent dense trajectory method in storage and performance.
- **Existing works**:
  - **Sampling strategy**: STIP, Cuboid detector, Dense trajectory (DT) etc.
  - **Descriptors**: HOG, HOF, HOG3D, MBH etc.
- **Our idea**:
  - To reduce the number of trajectory yet preserve the power of dense trajectory, we propose a motion boundary based dense sampling strategy, called DT-MB.
  - To enhance the discriminative power of DT, we propose a group of spatial temporal context descriptors, namely spatial co-occurrence HOG (S-CoHOG), S-CoHOF, S-CoMBH, temporal co-occurrence HOG (T-CoHOG), T-CoHOF and T-CoMBH.
- **Properties**:
  - **Faster**: the DT-MB deletes large number of points in sampling step, which sharply reduces the tracking cost for trajectories. It is faster than original DT.
  - **Discriminative**: spatial-temporal motion and appearance context information around pixels can deliver more complex motion and appearance structures.

Dense Trajectory on Motion Boundary

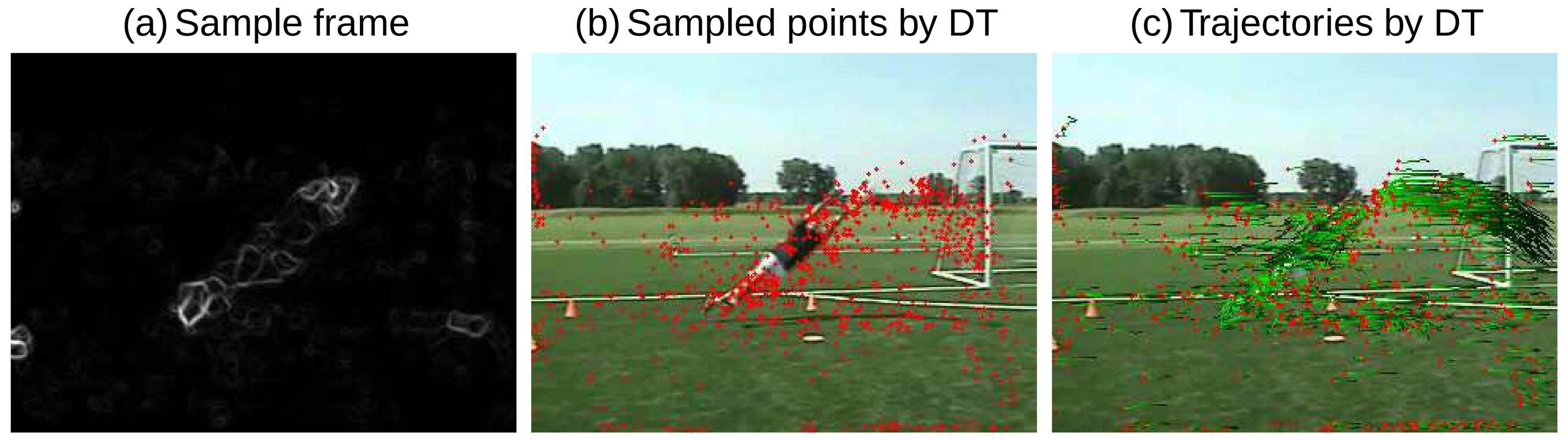
Motion Boundary: the maximum of gradient magnitudes between the horizontal and the vertical components of optical flow.

- **Implementation**:

1. **Sampling**: sample points in current frame on a grid by a step size w at S spatial scales. Two kinds of points will be removed: (1) the minimal eigenvalue of the covariance matrix of derivatives is less than a given threshold $T1$. (2) in the background of the binary mask estimated from motion boundary.
2. **Tracking and Filtering**: the same with Wang's [1].



(a) Sample frame (b) Sampled points by DT (c) Trajectories by DT



(d) Motion boundary image (e) Sampled points by DT-MB (f) Trajectories by DT-MB

Figure 1: Comparison of original DT and our Dense Trajectories on Motion Boundary.

Spatial-Temporal Context Descriptors

- **Spatial Context Descriptors**:

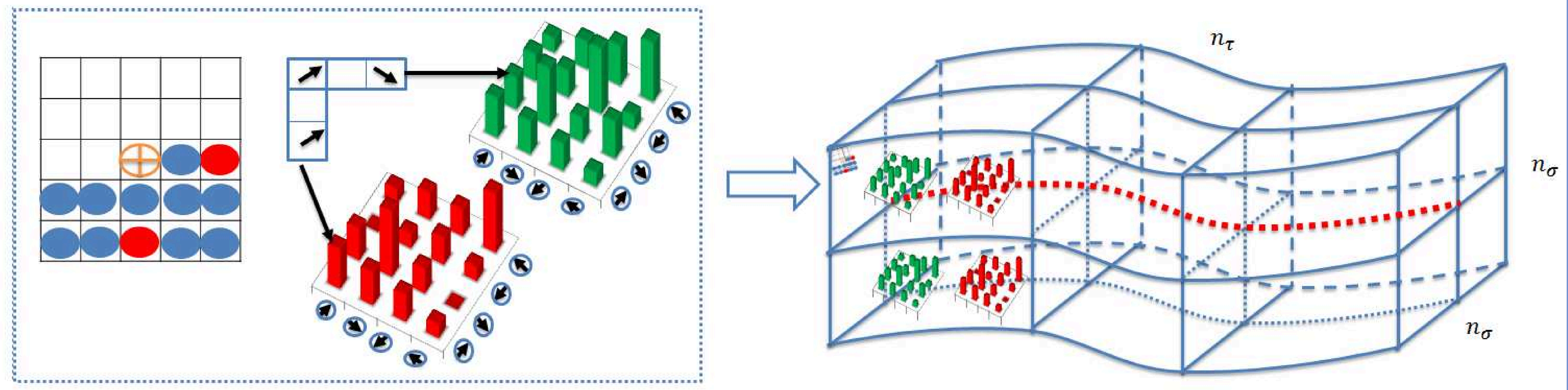
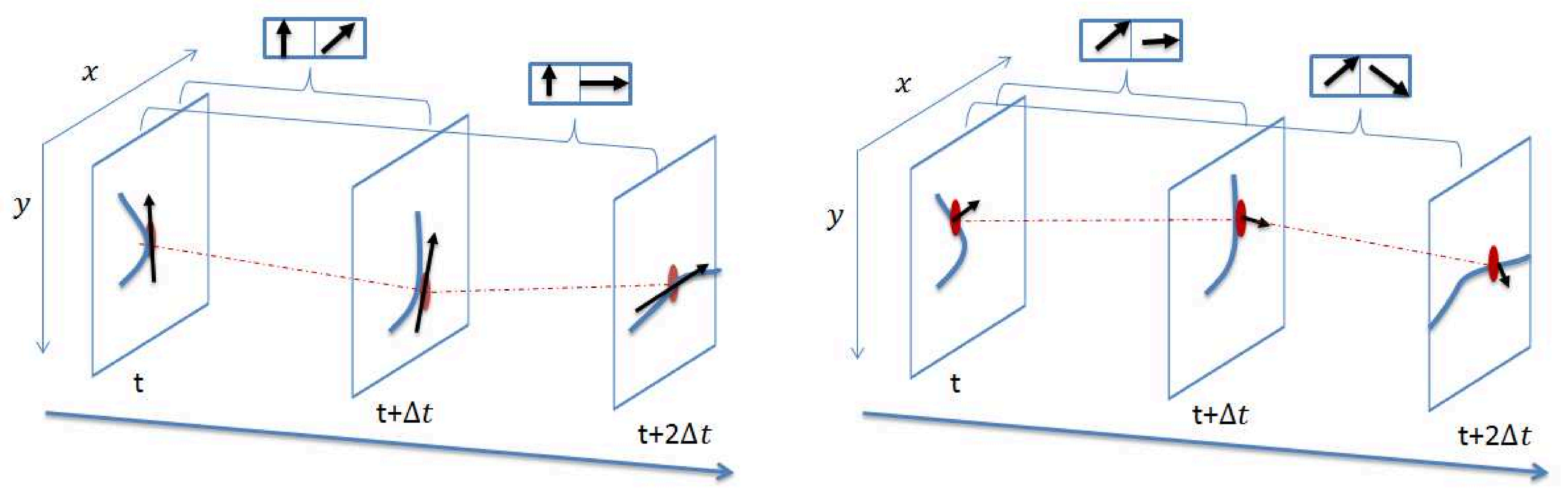


Figure 2: An example of spatial co-occurrence feature with grid of size $n_\sigma \times n_\sigma \times n_\tau$.

- **Co-occurrence matrix for offset (x,y) in a $m \times n$ patch:**

$$C_{x,y}(p,q) = \sum_{i=1}^{m}\sum_{j=1}^{n}\begin{cases}\frac{G(i,j)+G(i+x,j+y)}{2}, & \text{if } O(i,j)=p \text{ and } O(i+x,j+y)=q;\\ 0, & \text{otherwise.}\end{cases} \quad (1)$$

- **Temporal Context Descriptors**:



(a) T-CoHOG and T-CoMBH (b) T-CoHOF

Figure 3: Temporal co-occurrence descriptors. (a): pairs of gradient orientations in T-CoHOG or T-CoMBH. (b): pairs of optical flow orientations in T-CoHOF.

- **Representation**:
  - **Descriptor dimension**: $n_{bin} \times n_{bin} \times n_{offset} \times n_\sigma \times n_\sigma \times n_\tau$ for each type of S-Co ($n_{offset} = 2$) and T-Co ($n_{offset} = 1$) feature.
  - **Bag-of-features**: we use standard BoF pipeline to represent an action video.
  - **Classification**: LibSVM and one vs. all multi-channels RBF χ^2 kernel for multi-class classification.

References

1. Wang, Heng, Kläser, Alexander, Schmid, Cordelia and Liu, Cheng-Lin. Dense trajectories and motion boundary descriptors for action recognition, IJCV 2013.
2. Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. Learning realistic human actions from movies. In CVPR, 2008.
3. Liu, Jingen, Luo, Jiebo and Shah, Mubarak. Recognizing realistic actions from videos "in the wild". In CVPR, 2009.
4. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T. HMDB: A large video database for human motion recognition. In ICCV, 2011.
5. Le, Quoc V, Zou, Will Y., Yeung, Serena Y and Ng, Andrew Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In CVPR, 2011.
6. Sadanand, Sreemanananth and Corso, Jason J. Action bank: A high-level representation of activity in video. In CVPR, 2012.
7. Ji, Shuiwang, Xu, Wei and et al. 3D Convolutional Neural Networks for Human Action Recognition. PAMI, 2013.

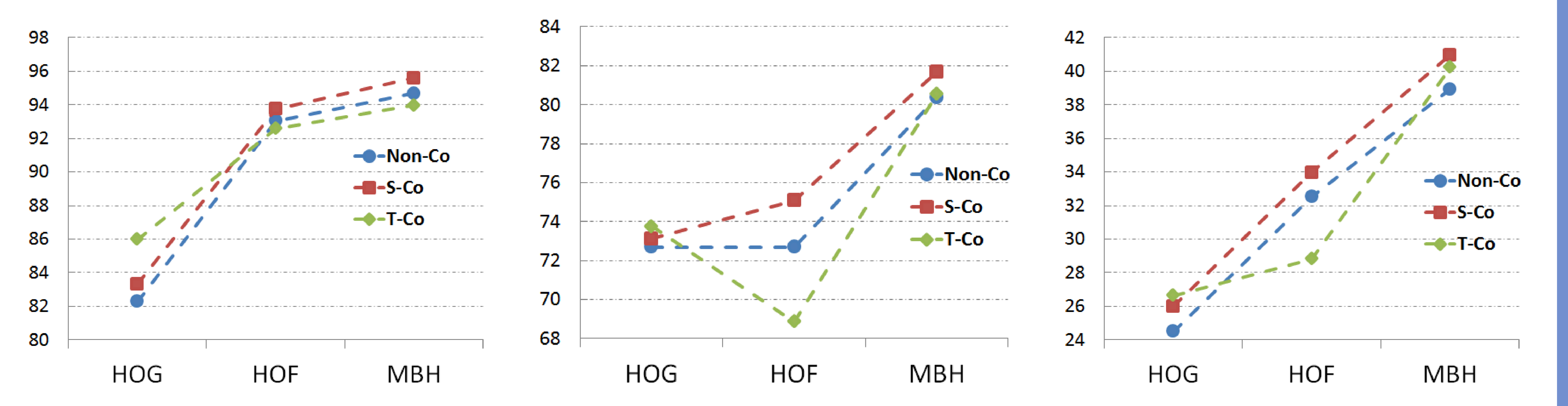
Experiment Results

- **Settings**: We conduct experiments on three datasets: KTH, UCF-Youtube, HMDB51.
- **DT-MB vs. DT**:

Table 1: Comparison of DT and DT-MB with all the raw DT descriptors.

Datasets		$T_{track}(ms)$	Trajectories/clip	fps	Accuracy (%)
HMDB51	DT	46.33	16,133	3.43	46.60
	DT-MB	**12.82**	4,512	4.63	46.03
YouTube	DT	39.01	37,542	4.71	84.25
	DT-MB	**6.60**	10,878	5.85	85.10
KTH	DT	11.72	2,185	12.85	94.81
	DT-MB	**4.00**	1,178	16.05	94.79

- **Spatio-temporal Context Descriptors**:



(a) KTH (b) YouTube (c) HMDB51

- **Descriptor combination**:

Table 2: Different combinations of descriptors using standard BOF.

Combination	KTH	YouTube	HMDB51
Trajectory+HOG+HOF+MBH	93.63	84.25	45.90
HOG+HOF+MBH	93.98	83.48	45.88
Trajectory+S-Co + T-Co	94.79	85.70	48.98
S-Co + T-Co	94.21	85.33	48.89
All combined	94.10	86.30	49.22
Best combined	95.60	86.56	49.22

- **Comparison**:

Table 3: Compare our results to the state-of-the-art results.

KTH		YouTube		HMDB51	
Laptev *et al.*[2]	91.8	Liu *et al.*[3]	71.2	Kuehne *et al.*[4]	23
Le *et al.*[5]	93.9	Le *et al.*[5]	75.8	Sadanand *et al.*[6]	26.9
Ji *et al.*[7]	90.2	B. *et al.*[8]	76.5	Orit *et al.*[9]	29.2
Wang *et al.*[1]	95	Wang *et al.*[1]	84.1	Wang *et al.*[1]	46.6
Our Method	**95.6**	Our Method	**86.56**	Our Method	**49.22**

8. Bhattacharya, Subhabrata and Sukthankar, Rahul. A probabilistic representation for efficient large scale visual recognition tasks. In CVPR, 2011.
9. Kliper-Gross, Orit and Gurovich, Yaron. Motion Interchange Patterns for Action Recognition in Unconstrained Videos. In ECCV, 2012.