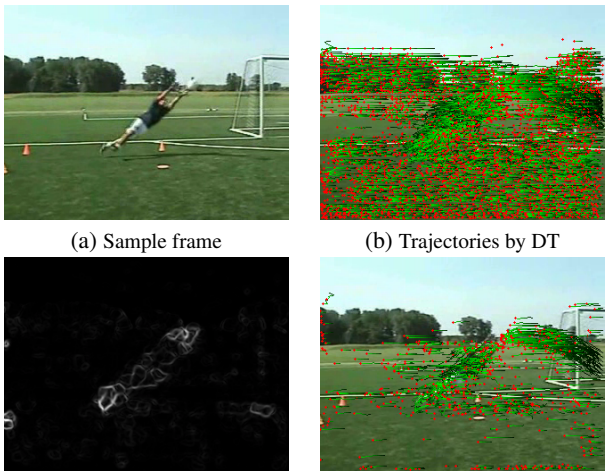
Exploring Motion Boundary based Sampling and Spatial-Temporal Context Descriptors for Action Recognition

Xiaojiang Peng[1,2]
xiaojiangp@gmail.com
Yu Qiao[2,3]
yu.qiao@siat.ac.cn
Qiang Peng[1]
qpeng@home.swjtu.edu.cn
Xianbiao Qi[2]
qixiaobiao@gmail.com

[1] Southwest Jiaotong University, Chengdu, P.R. China
[2] Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] The Chinese University of Hong Kong



(a) Sample frame (b) Trajectories by DT

(c) Motion boundary image (d) Trajectories by DT-MB

Figure 1: The trajectories captured by original DT and our DT-MB.

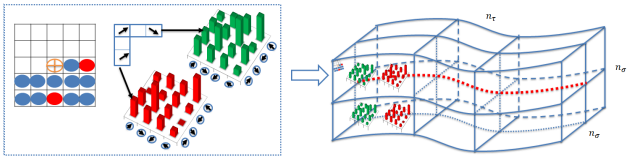
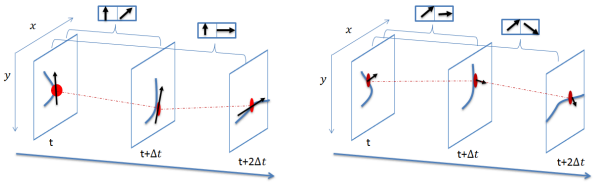


Figure 2: An example of spatial co-occurrence features with grid of size $n_\sigma \times n_\sigma \times n_\tau$.



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

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

The most important problem in action recognition is how to represent an action video. The approaches can be roughly divided into four categories: (1) human pose based approaches which utilize human structure information; (2) global action template based approaches which capture appearance and motion information on the whole motion body; (3) local feature based approaches which mainly extract valid space-time cuboids; (4) unsupervised feature learning based methods which learn the representation by hierarchical networks. Among these approaches, local feature with bag-of-features (BoF) framework is perhaps the most popular way for action recognition.

With the mentioned popular pipeline, Wang et al. [2] proposed dense trajectory (DT) based features for action video representation and achieved state-of-the-art performance on several action datasets recently. Though its great power, the DT method is expensive in memory storage and computation due to the large number of dense sampled points. In this paper, we improve the DT method in two folds. Firstly, we introduce a motion boundary based dense sampling strategy, called *DT-MB*, which greatly reduces the number of valid trajectories while preserves the discriminative power. Secondly, we develop a set of co-occurrence descriptors which describe the *spatial-temporal context* of motion trajectories.

Our DT-MB is partly implied by MBH descriptor [2] and motion boundary contour system (BCS) in neural dynamics of motion perception [1]. It constrains the sampled points on large magnitude regions of motion boundary image in the sampling step. A comparison with original DT method is illustrated in Fig.1. Our sampling approach removes a large number of points which are not on the motion foreground.

We propose spatial-temporal co-occurrence HOG, HOF and MBH to further enhance the performance of DT. The pipeline of spatial co-occurrence feature in a regularized spatial-temporal grid is shown in Fig.2, and the temporal one is depicted in Fig.3. The spatial co-occurrence HOG [3], HOF and MBH aim to capture complex spatial structures of appearance and motion. Our novel temporal co-occurrence descriptors depict clear motion and appearance changes from successive patches.

Our results of individual co-occurrence descriptors on three datasets are illustrated in Fig.4. It indicates that temporal context information for pure spatial feature is more effective, and spatial context information for pure temporal features is beneficial. Table 1 shows the combined results in detail.

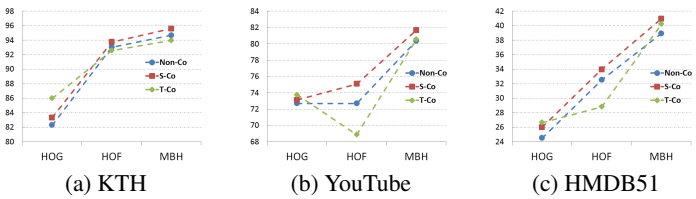
Table 1: Different combinations of descriptors using standard BOF.

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

[1] Stephen Grossberg and Ennio Mingolla. Neural dynamics of motion perception: direction fields, apertures, and resonant grouping. *Perception & psychophysics*, 53(3):243–278, 1993.

[2] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, pages 1–20, 2012.

[3] Tomoki Watanabe, Satoshi Ito, and Kentaro Yokoi. Co-occurrence histograms of oriented gradients for pedestrian detection. *Advances in Image and Video Technology*, pages 37–47, 2009.



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

Figure 4: Percentage accuracies of all the individual descriptors on three datasets. "Non-Co" corresponds to the original descriptors in [2].