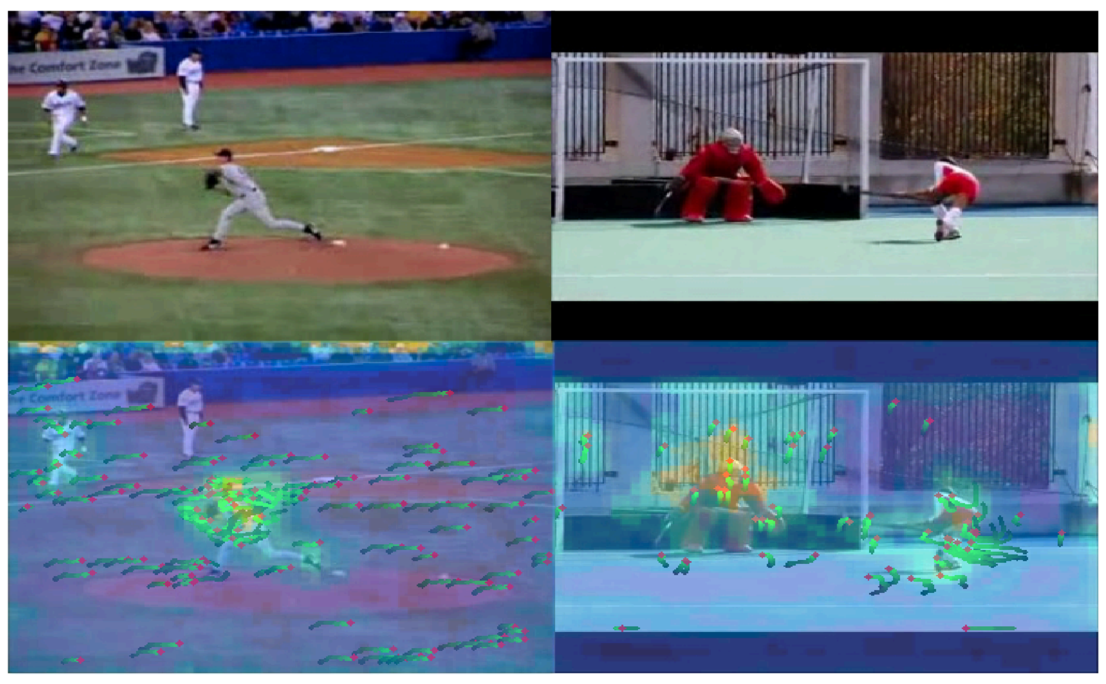
Improved Dense Trajectory with Cross Streams

Katsunori Ohnishi, Masatoshi Hidaka, Tatsuya Harada
The University of Tokyo

ACM multimedia 2016

Introduction:



Visualized iDT and feature map
from convolutional layer in temporal net.

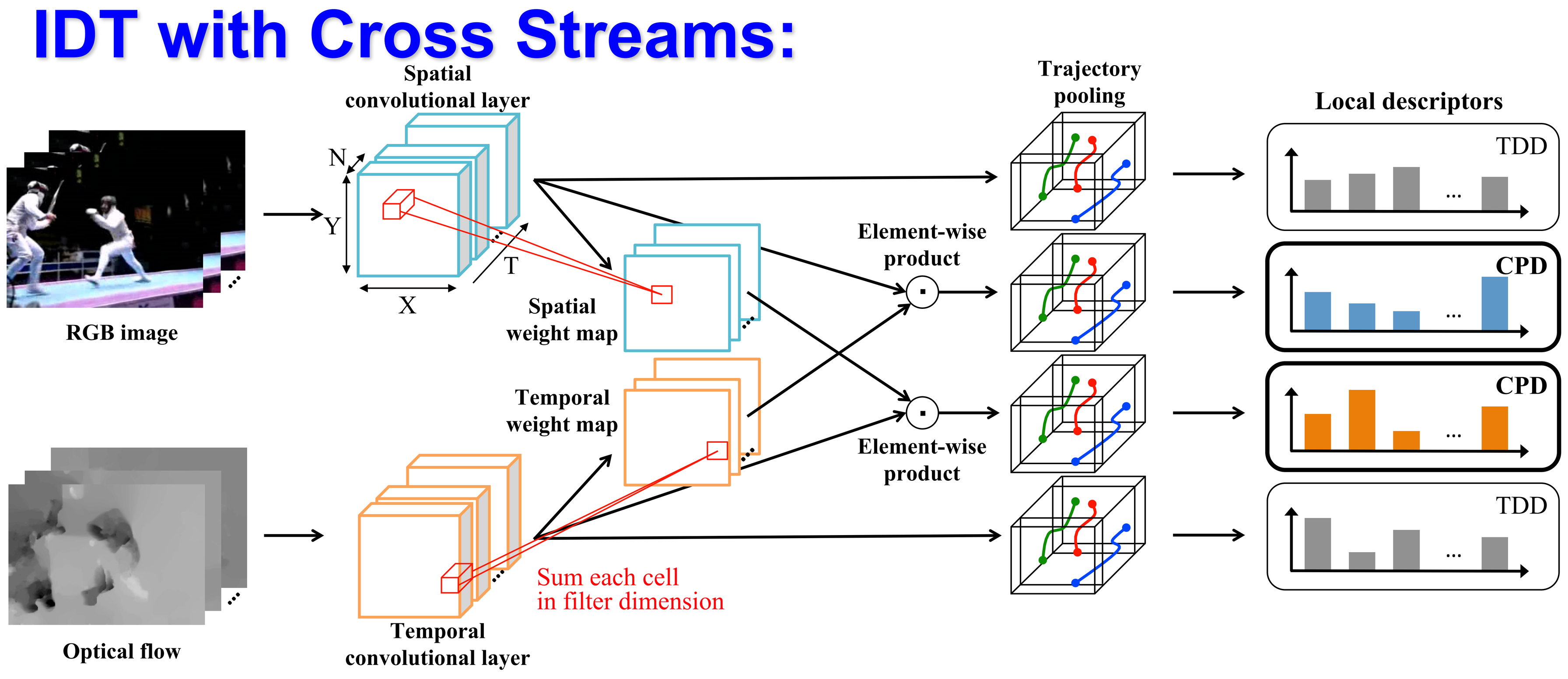
- Topic: action recognition
- Existing works:
  - iDT [H. Wang, et al., ICCV13] removes dense trajectories in background images considering camera motion.
  - Two-stream [K. Simonyan, et al., NIPS14] separately learns two CNNs, spatial net with RGB and temporal net with optical flow.
  - TDD [L. Wang, et al., CVPR15] combines iDT and Two-stream.

Two main shortcomings in existing works

1. iDT cannot completely remove the background trajectories from videos captured by a shaking camera
2. Separate CNN learning sometimes lacks other important information that can be obtained only when spatial and temporal information are combined together

Goal:
Design a new descriptor that contains complementary information between spatial and temporal networks for action recognition

IDT with Cross Streams:



- TDD [L. Wang, et al., CVPR15]

$$TDD(P^k, \tilde{C}_b^a) = \sum_{l=1}^{L} \tilde{C}_b^a((r_x \times x_l^k), (r_y \times y_l^k), t_l^k)$$

$$\tilde{C}_{st}(x,y,n,t) = C(x,y,n,t) / \max_{x,y,t} C(x,y,n,t)$$

$$\tilde{C}_{ch}(x,y,n,t) = C(x,y,n,t) / \max_{n} C(x,y,n,t)$$

$$(r_x, r_y) = (X/V_W, Y/V_h) \qquad \text{Trajectory point: } (x_l^k, y_l^k, t_l^k)$$

$$a \in \{sp, tmp\} \qquad b \in \{st, ch\} \qquad C \in \mathbb{R}^{X \times Y \times N \times T}$$

Instead of originally pooled features (HOG, HOF, and MBH), TDD pools normalized convolutional layers along iDT.

- Cross-stream pooled descriptors (CPD)

$$CPD(P^k, \tilde{C}_b^a, W_b^{\bar{a}}) = \sum_{l=1}^{L} W_b^{\bar{a}}(x_l^k, y_l^k, t_l^k) \times \tilde{C}_b^a((r_x \times x_l^k), (r_y \times y_l^k), t_l^k)$$

$$W_b^a(x,y,t) = \sum_{n=1}^{N} \tilde{C}_b^a(x,y,n,t)$$

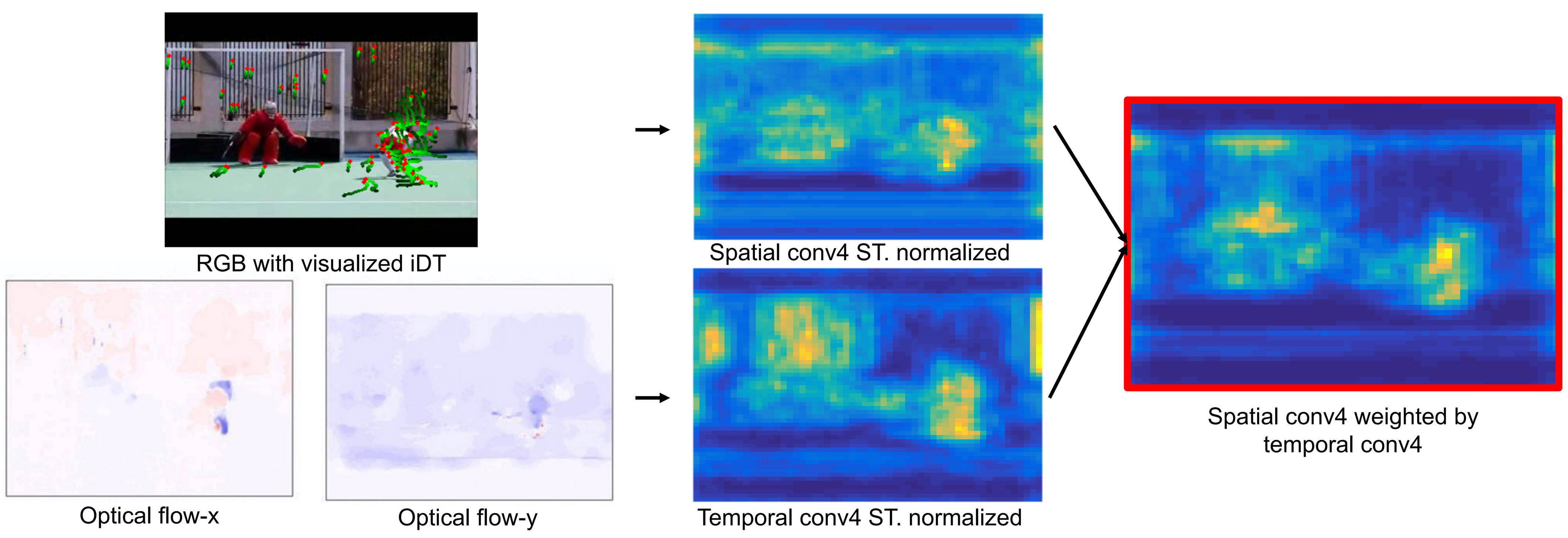
CPD multiplies spatial and temporal convolutional layers element-wise and pools the resulting four-dimensional matrix along iDT.

- In order to enhance motion-important regions in a spatial convolutional layer and appearance-important regions in a temporal convolutional layer.

Experiment:

- Dataset
  - UCF101
    101 classes, 13k videos
  - HMDB51
    51classes, 6.8k videos

Sum of filter activations



Mean accuracy of CPD and other baseline methods on HMDB51 and UCF101

*1: L. Wang, et al., arXiv:1507.02159, 2015.

Algorithm	HMDB51	UCF101
iDT & FV	57.2%	85.9%
Two-stream	59.4%	88.0%
TDD & FV	63.2%	90.3%
Two stream (VGG16)	61.9%	91.4% *1
Spatial net (VGG16 w/o flip&crop)	39.7%	75.5%
Temporal net (VGG16 w/o flip&crop)	53.6%	81.0%
Two stream (VGG16 w/o flip&crop)	59.3%	87.6%
TDD (VGG16) & FV	63.2%	91.3%
TDD (VGG16) & VLAD	65.0%	92.0%
CPD & VLAD **(ours)**	65.2%	91.8%
TDD (VGG16) & VLAD + CPD & VLAD **(ours)**	**66.2%**	**92.3%**

Comparison with the state-of-the-art methods

HMDB51		UCF101	
iDT & FV	57.2%	iDT & FV	85.9%
iDT & stacked FV [X. Peng, et al., ECCV14]	56.2%	C3D [D. Tran, et al., ICCV15]	85.2%
+ iDT & FV	*(66.8%)*	+iDT & FV	*(90.4%)*
$F_{ST}CN$ [L. Sun, et al., ICCV15]	59.1%	$F_{ST}CN$	88.1%
LATE [C. Feichtenhofer, et al., CVPR15]	62.2%	MIFS	89.1%
TDD & FV	63.2%	TDD & FV	90.3%
+iDT & FV	*(65.9%)*	+ iDT & FV	*(91.5%)*
Video darwin [B. Fernando, et al., CVPR15]	63.7%	Hybrid LSTM [Z. Wu, et al., ACMMM15]	91.3%
MIFS [Z. Lan, et al., CVPR15]	65.1%	Two stream (VGG16)	91.4%
CPD **(ours)**	65.2%	CPD **(ours)**	91.8%
TDD + CPD **(ours)**	**66.2%**	TDD + CPD **(ours)**	**92.3%**