BMVC 2016 poster presentation

Recognition of Transitional Action for Short-Term Action Prediction using Discriminative Temporal CNN Feature

Hirokatsu Kataoka[1], Yudai Miyashita[2], Masaki Hayashi[3,4], Kenji Iwata[1], Yutaka Satoh[1]

[1]AIST, Japan [2]Tokyo Denki Univ., Japan [3]Liquid Inc., Japan [4]Keio Univ., Japan

Motivation

- Goal
  - Accurate "short-term action prediction"
- Problems in action analysis
  - Recognition is NOT predictable
  - Prediction is NOT reliable
- Applications
  - Active safety, autonomous driving
  - Robots

Framework	Problem
Action Recognition	$f(F^A_{1...t}) \rightarrow A_t$
Early Action Recognition	$f(F^A_{1...t-L}) \rightarrow A_t$
Action Prediction	$f(F^A_{1...t}) \rightarrow A_{t+L}$
Transitional Action Recognition	$f(F^{TA}_{1...t}) \rightarrow A_{t+L}$

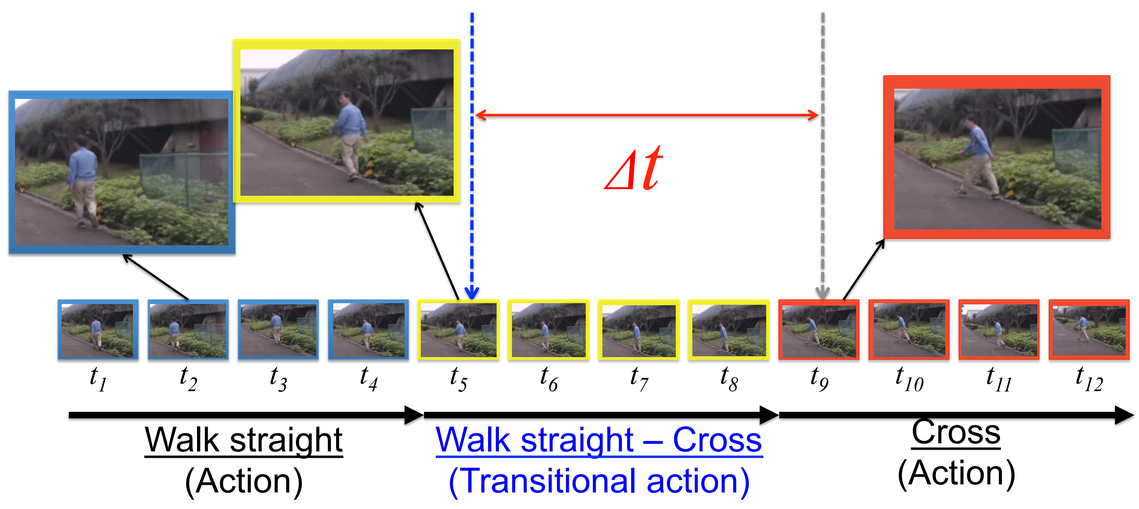
Framework

- IDEA
  - Action-class while an action is transitive (see below)
- Contributions
  - Definition of transitional action for short prediction
  - Subtle Motion Descriptor (SMD) to classify TA*1 and NA*2

*1: Transitional action *2: Normal action

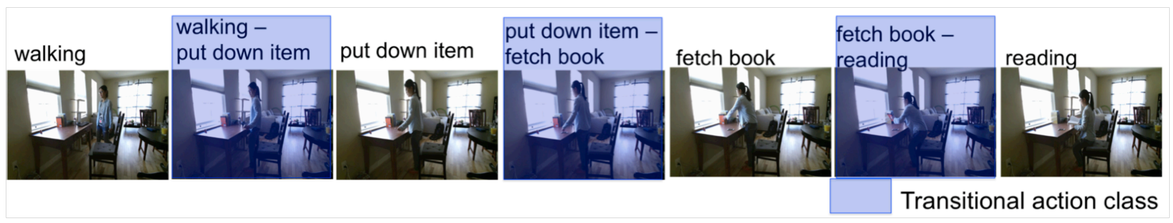
【Proposal】
Short-term action prediction
recognize "cross" at time t_5

【Previous works】
Early action recognition
recognize "cross" at time t_9



Transitional Action?

- Transitional action is defined as the transition class between actions (see below)
  - TA: "walking – put down item" between NA: "walking" and NA: "put down item"
  - The TA classes and NA classes are partially overlapped each other
    - But no more than 5 frames overlap

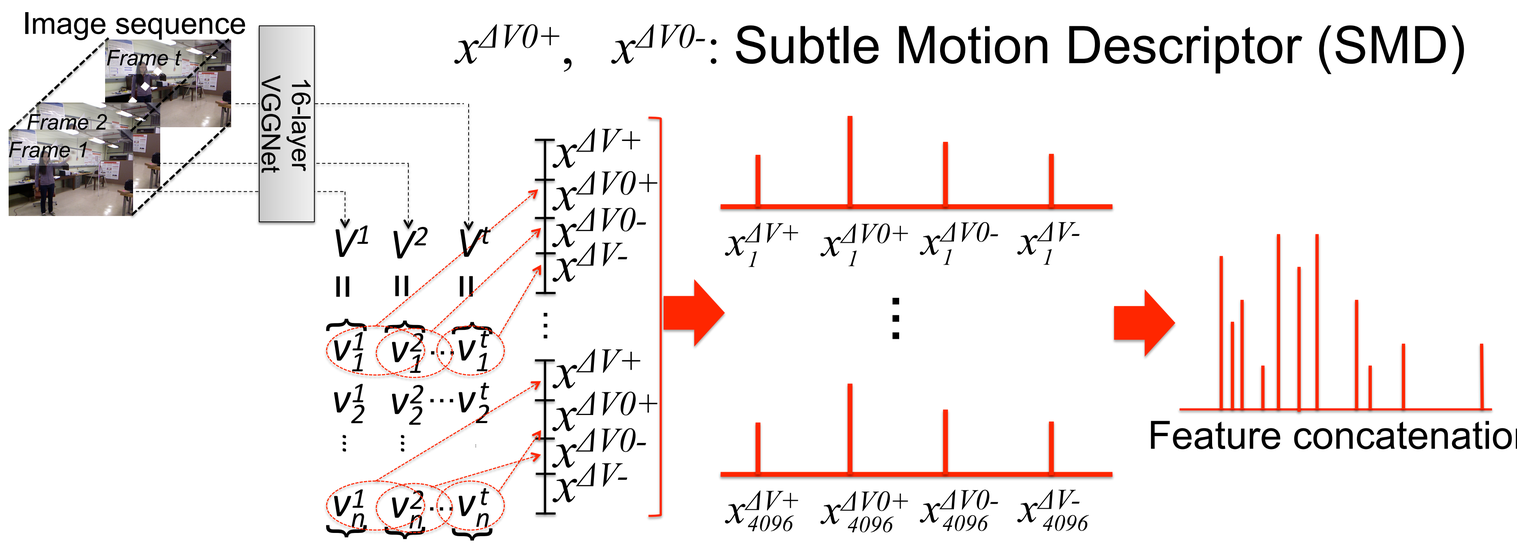


Subtle Motion Descriptor (SMD)

- SMD is used to identify the sensitive differences between TA and NA
  - The SMD is based on pooled time series (PoT) [Ryoo+, CVPR15]
  - Temporal pooling with zero-around elements ($x^{\Delta V0-}$, $x^{\Delta V0+}$)

$$x_i^{\Delta V^+} = \Sigma_{t=t_s}^{t_e} h_i^+(t), \quad x_i^{\Delta V^{0+}} = \Sigma_{t=t_s}^{t_e} h_i^{0^+}(t)$$
$$x_i^{\Delta V^-} = \Sigma_{t=t_s}^{t_e} h_i^-(t), \quad x_i^{\Delta V^{0-}} = \Sigma_{t=t_s}^{t_e} h_i^{0^-}(t)$$

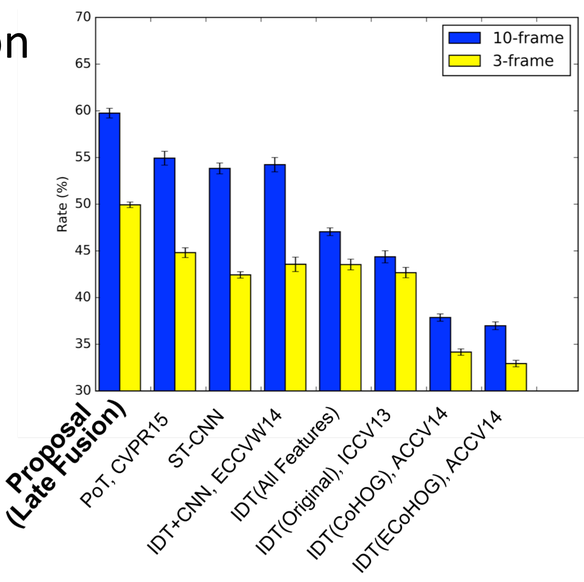
$$\begin{cases} h_i^+(t) = |\Delta v_i^t| \ \ (\Delta v_i^t > TH) \\ h_i^{0^+}(t) = |\Delta v_i^t| \ \ (0 < \Delta v_i^t < TH) \\ h_i^{0^-}(t) = |\Delta v_i^t| \ \ (-TH < \Delta v_i^t < 0) \\ h_i^-(t) = |\Delta v_i^t| \ \ (\Delta v_i^t < -TH) \end{cases}$$



Experiments

- NTSEL (Traffic; TA1, NA3), UTKinect-Action (Indoor; TA8, NA10), Watch-n-Patch (Indoor; TA10, NA10)
  - Threshold? – good for 0.03 ~ 0.05
  - Frame accumulation? – 10-frame for state-of-the-art, 3-frame for faster prediction
  - FC layer – Layer 6 is better

	% on NTSEL		% on UT		% on WnP	
	10 frm	3 frm	10 frm	3 frm	10 frm	3 frm
Ours	**99.18**	**85.78**	**99.19**	**69.77**	**59.75**	**49.93**
PoT, CVPR15	97.00	77.15	92.00	65.46	54.93	44.81



Conclusion

- Definition of transitional action (TA) for short-term prediction
  - The TA allows us to produce earlier and accurate prediction
- Proposal of subtle motion descriptor (SMD)
  - Outstanding results with 3-frame feature accumulation

Reference

[1] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.