3D Convolutional Neural Network with Multi-Model Framework for Action Recognition

Longlong Jing*, Yuancheng Ye*, Xiaodong Yang^, Yingli Tian*

*The City University of New York, NY, USA

^NVIDIA Research, CA, USA

Motivation

- Home Security

Motivation

- Public Security & Service
  - public agency
  - financial service
  - manufacturer
  - retailer

- Intelligent Analysis
  - big data
  - trends out of data

WHEREVER YOU GO, WHATEVER YOU DO, WHOEVER YOU ARE,

YOU ARE UNDER SURVEILLANCE

BECAUSE YOU ARE A POTENTIAL CRIMINAL. PERHAPS YOU SECRETLY DOUBT THE SANCTITY OF CORPORATE PROPERTY, OR THE VALIDITY OF LAWS MADE BY THE RICH TO GOVERN THE POOR, OR THE SOUNDNESS OF CAPITALISM ITSELF—WE CAN'T AFFORD TO ASSUME YOU DON'T. THAT'S WHY THERE ARE VIDEO CAMERAS POINTED AT EVERY CASHIER AND POLICE CARS CIRCLING EVERY BLOCK. LEFT TO ITSELF, A STATE OF DISORDER AND INEQUITY RETURNS TO EQUILIBRIUM; OUR JOB IS TO PERPETUATE THIS ONE INDEFINITELY.

DEPARTMENT OF HOMELAND SECURITY

"in suspicion we trust!"

www.crimethinc.com/supervision

Action Recognition

- Identifying the activity people doing in videos.
- Capturing both the spatial and temporal information of the activity.

Playing basketball

Doing push-up

Playing baseball

Existing Algorithms

- **Hand-designed feature-based**

  Extracting features by hand-designed algorithms.

  Applying classifiers to the extracted features.
- **2D CNN-based**

  Treating a video as a set of frames.

  Applying 2D-CNN in each frame.
- **3D CNN-based**

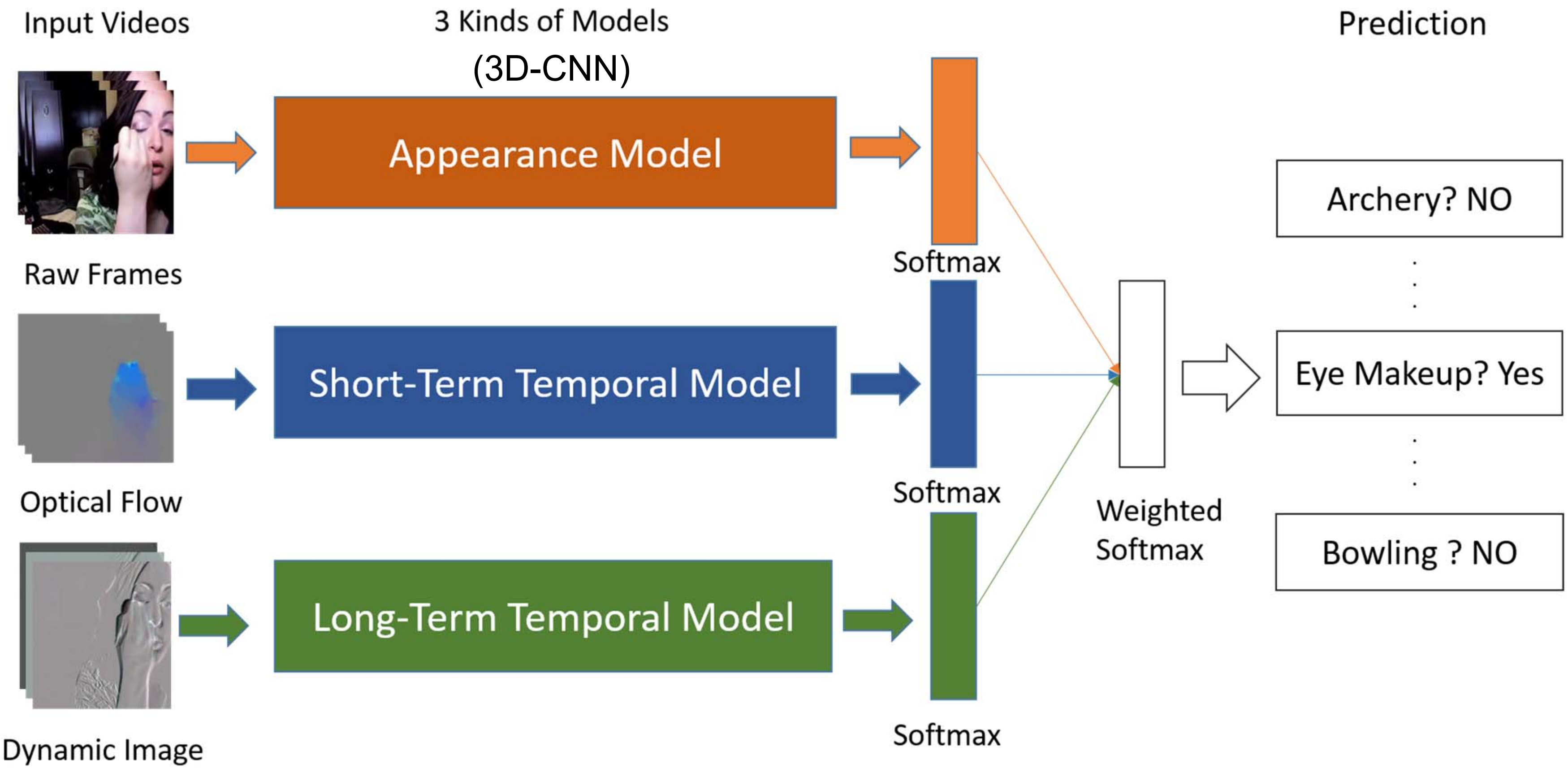
  Dividing each video into small clips with fixed length.

  Applying 3D-CNN in each clip.
- **LSTM-based**

  Treating the whole video as frame sequence.

  Handle videos with variable lengths.

Our Framework



Appearance Model

- The network is an 11-layer 3D-CNN.
- The input of the appearance model is 16 consecutive RGB frames.
- The network captures the appearance information from these clips.

Short-Term Temporal Model

- The network is an 11-layer 3D-CNN.

- The input is 16 consecutive optical flow images. Optical flow is calculated by the method [Brox et al. 2004].

- Optical flow captures the motion between frames and boundary of moving objects.

T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” ECCV’04.

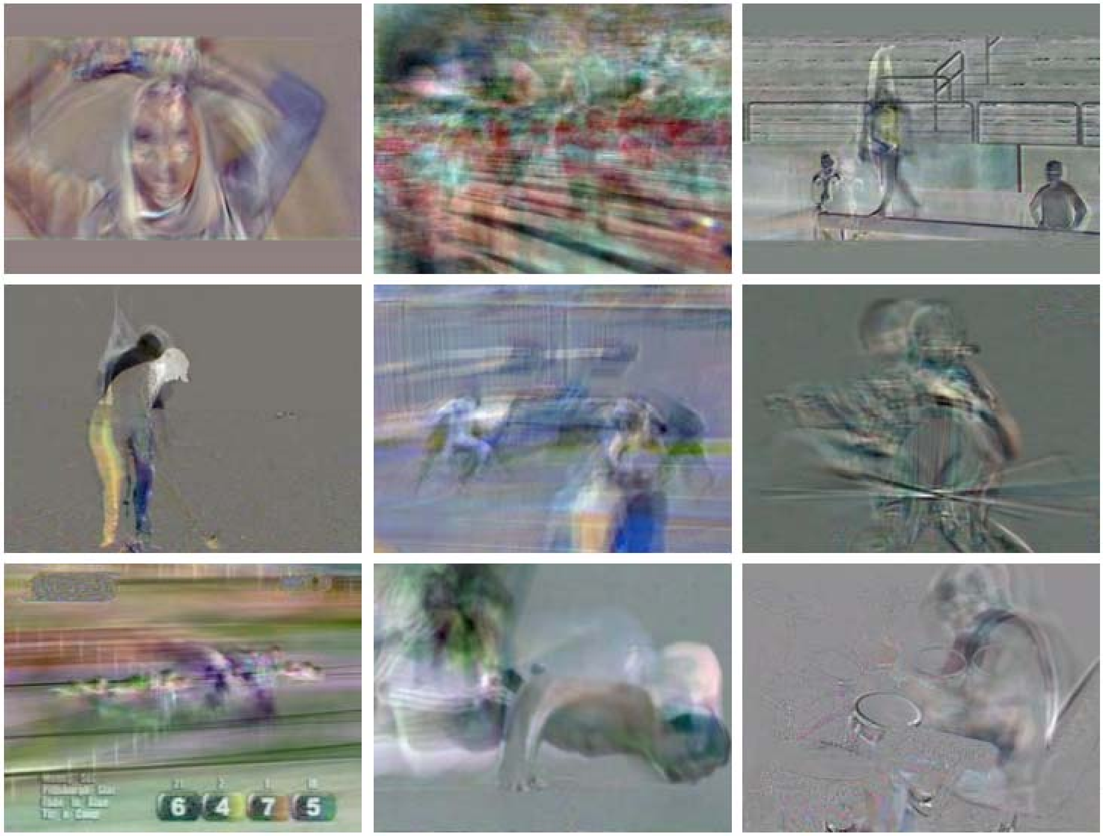
Long-Term Temporal Model

- The network is an 11-layer 3D-CNN.

- The input of the long-term temporal model is 16 dynamic images (generated from the whole video).

- The network captures the long-term temporal information from the whole video.

Dynamic Image Examples



The figure is from [Bilen et al.]

Fernando et al., “Modeling video evolution for action recognition,” CVPR 2015.
Bilen et al., “Dynamic image networks for action recognition,” CVPR 2016.

Dynamic Image

- Applied a ranking machine with approximate rank pooling.

- Directly applying approximate rank pooling on the raw image pixels of a video.

- The parameters of the frames can be pre-computed, which makes the computation of dynamic image very efficient.

Dynamic Image Generation

- $\rho(I_1, \ldots, I_T, \varphi) = \sum_{t=1}^{T} \alpha_t \varphi(I_t)$. --- approximate rank pooling

- $\alpha_t = 2(T - t + 1) - (T + 1)(H_T - H_{t-1})$.

- $H_t = \sum_{t=1}^{t} 1/t$. -- the *t-th* Harmonic number

- $\varphi(I_t)$ is the pixels for the t^{th} frame, T is the length of the video.

Dynamic Image actually is the linear combination of the frames.

Bilen et al., “Dynamic image networks for action recognition,” CVPR 2016.

Datasets

- UCF101

  13K videos (10K training, 3K testing).

  101 categories.

  Frame rate: 30FPS.

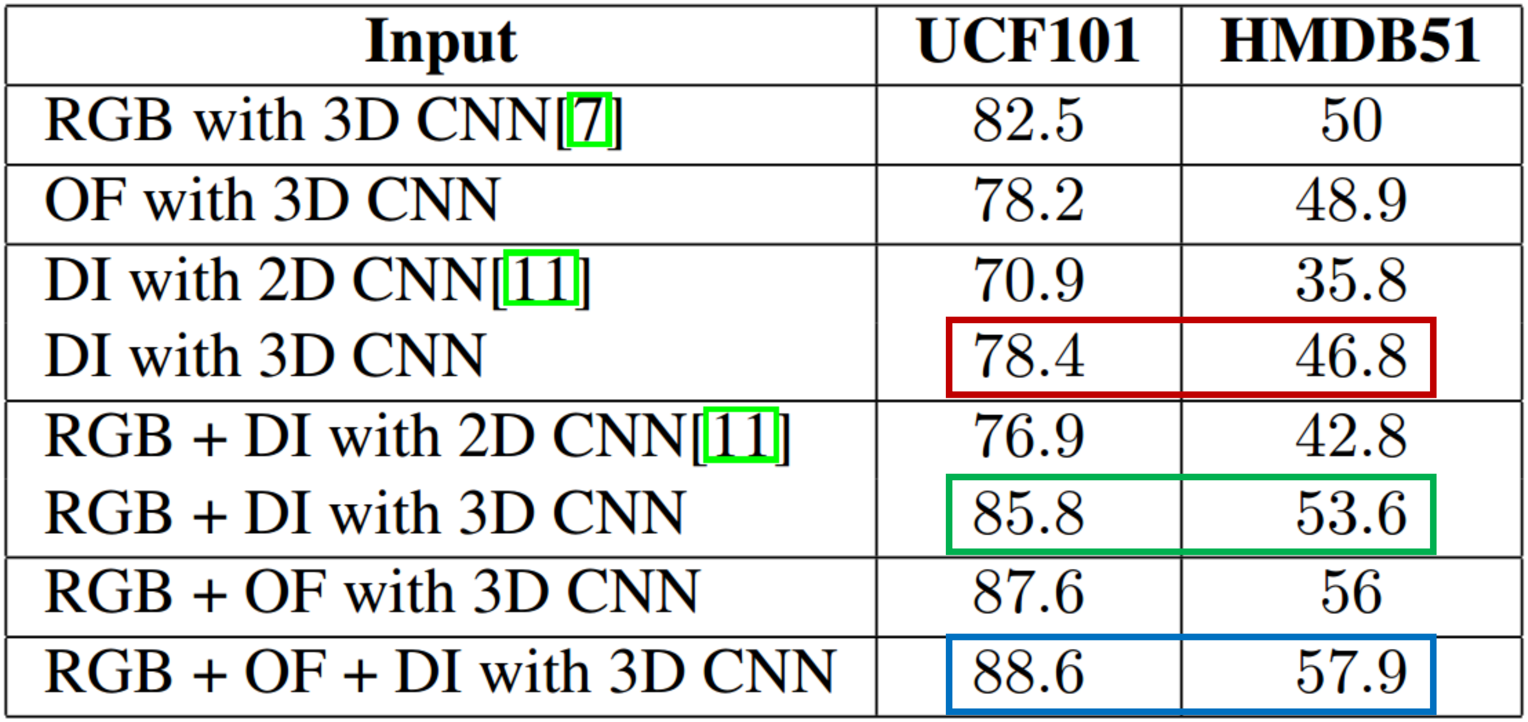
- HMDB51

  7K videos (5K training, 2K testing).

  51 categories.

  Frame rate: 25FPS.

Experimental Results



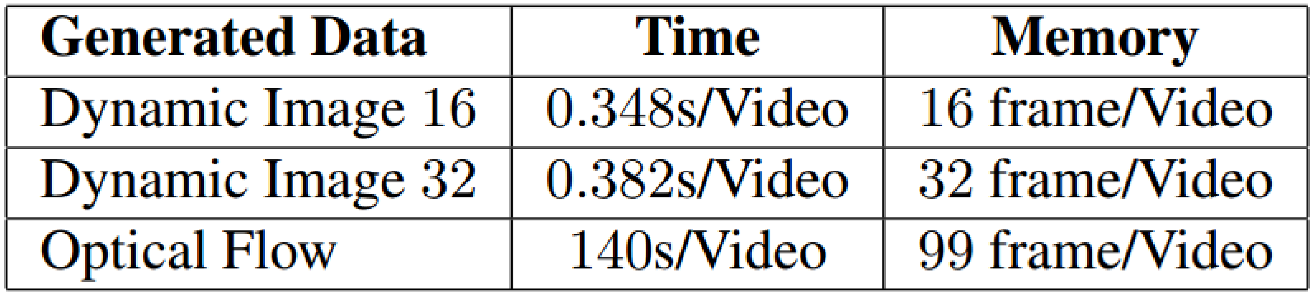
Input	UCF101	HMDB51
RGB with 3D CNN[7]	82.5	50
OF with 3D CNN	78.2	48.9
DI with 2D CNN[11]	70.9	35.8
DI with 3D CNN	78.4	46.8
RGB + DI with 2D CNN[11]	76.9	42.8
RGB + DI with 3D CNN	85.8	53.6
RGB + OF with 3D CNN	87.6	56
RGB + OF + DI with 3D CNN	88.6	57.9

[7] Tran et al. Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015.
[11] Bilen et al. Dynamic Image Networks for Action Recognition, CVPR 2016.

Efficiency of Dynamic Image

- The computation of dynamic image is very efficient.

- The computation of dynamic Image needs less memory than others forms.



Generated Data	Time	Memory
Dynamic Image 16	0.348s/Video	16 frame/Video
Dynamic Image 32	0.382s/Video	32 frame/Video
Optical Flow	140s/Video	99 frame/Video

Summary

- We proposed a new framework for action recognition by combining multiple feature models.

- We compressed a video into 16 frame dynamic images. The dynamic image preserves the overall temporal information.

- The computation of dynamic image is very efficient for real-time applications.

Thank You & Questions!