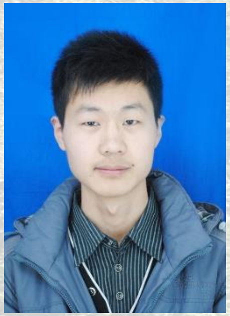
中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

模式识别国家重点实验室
National Laboratory of Pattern Recognition

Cascaded Temporal Spatial Features for Video Action Recognition

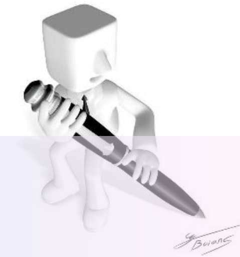
Tingzhao Yu[1;2], Huxiang Gu[1], Lingfeng Wang[1], Shiming Xiang[1], Chunhong Pan[1]





Code Available at *https://github.com/Tsingzao/motion_image*

1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
2. School of Computer and Control Engineering, University of Chinese Academy of Sciences



Challenges and Datasets

Thumos

Charades

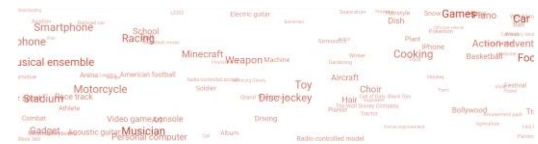
Youtube-8M

ODAR

LSVC

ActivityNet

Youtube-8M

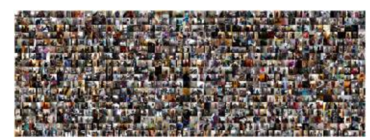


ActivityNet

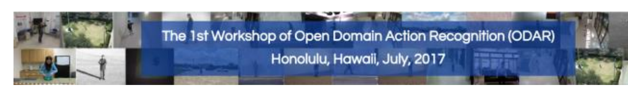
FCVID

UCF

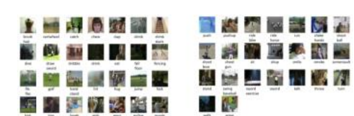
Charades



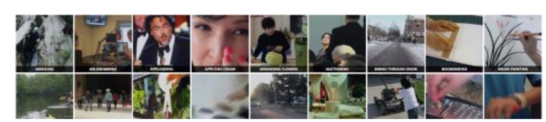
ODAR



HMDB



Kinects



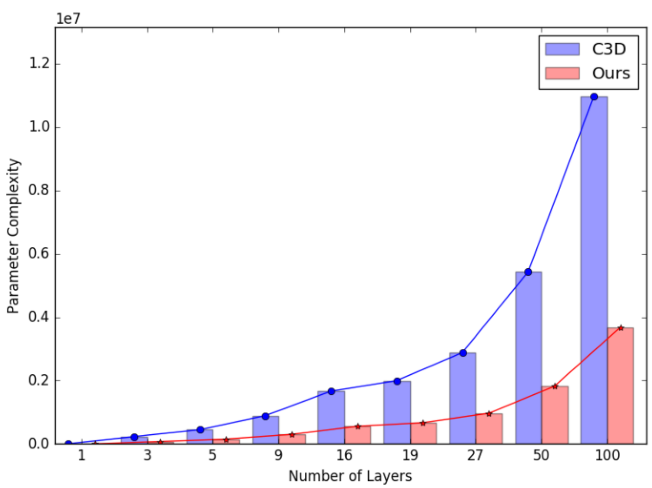
Video based Action Recognition

- **Problem Formulation:**
  - **Recognize the actions being taken place, i.e., video sequence**
- **Assumption:**
  - **Known action classes – Video Classification (trimmed video)**

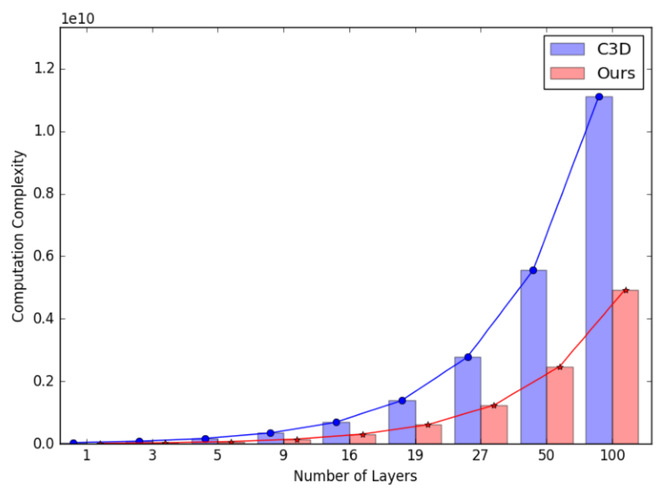
Handcrafted Features
- 3D-MHI [ICCV2009]
- STIP [IJCV2005]
- DT&iDT [CVPR2011]
- DenseSIFT [AN 2013]
- 3D-HoG [BMVC2008]

Deep Features
- Frame by Frame[CVPR 2014]
- 3D Convolution [ICCV2015]
- Two Stream [NIPS2016]
- Dynamic&Factorize [CVPR]
- LSTM & RNN [NIPS CVPR]

Cascaded Temporal Spatial Feature



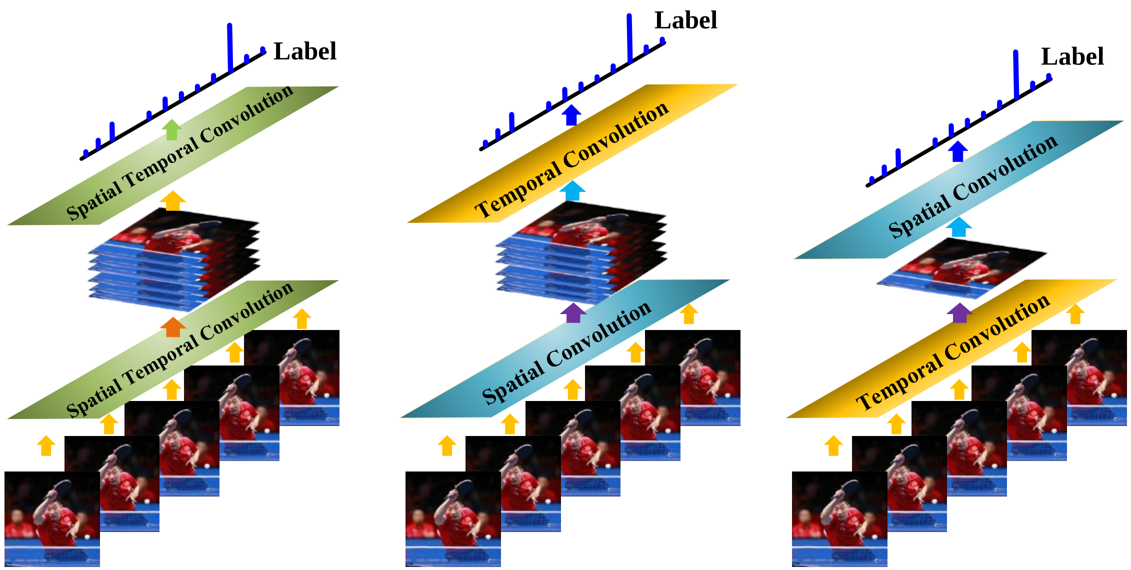
(a)



(b)

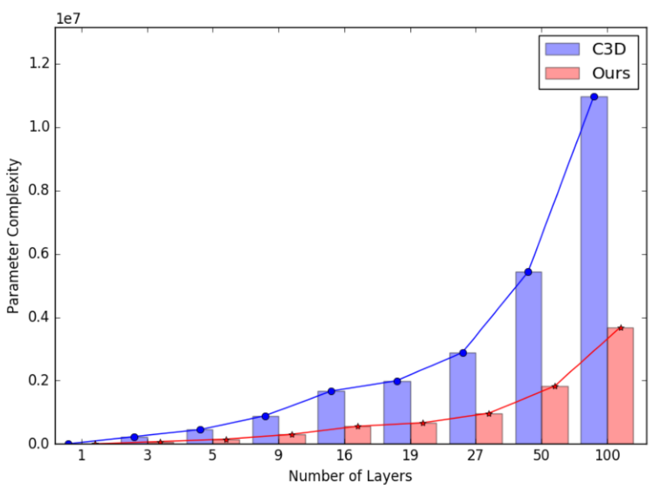
The ***motivation*** behind this design is to achieve deep nonlinear feature representations with reduced network parameters.

Cascaded Temporal Spatial Feature

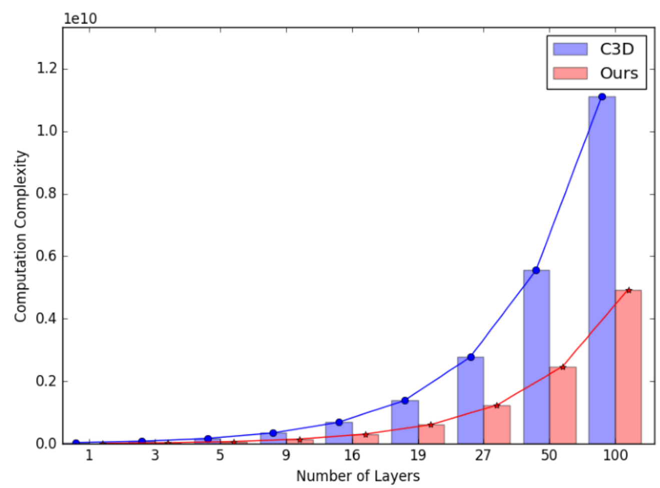


(a) (b) (c)

- (a) C3D exploits the spatial-temporal features simultaneously
- (b) F_{ST}CN factorizes the spatial-temporal features into spatial and temporal domain
- (c) the Proposed Architecture decouples the spatial-temporal features into cascaded temporal and spatial domain.



(a)

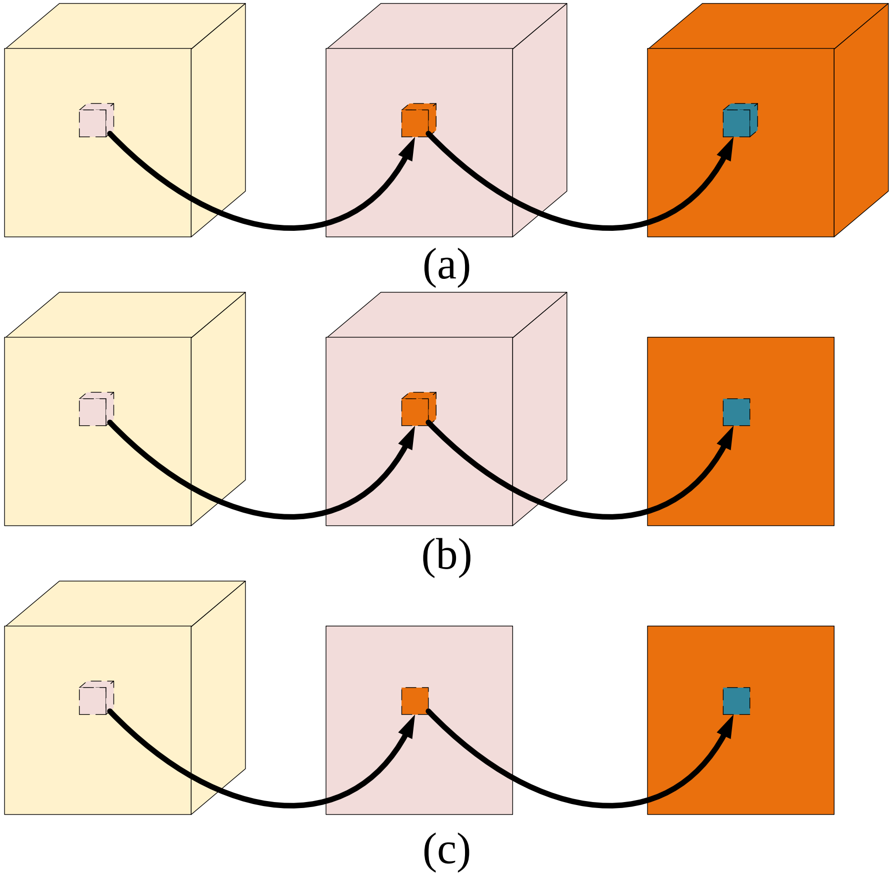


(b)

The ***motivation*** behind this design is to achieve deep nonlinear feature representations with reduced network parameters.

Analysis

- **From 3D Convolution to Decoupled 1D and 2D**



3D Convolution [2]

2D * 1D Convolution [7]

1D * 2D Convolution

Architecture

- **From 3D Convolution to Decoupled 1D and 2D**

$$v^{xy} = \sum_{m} \sum_{p=1}^{P} \sum_{q=1}^{Q} K^{pq} \; v^{(x+p)(y+q)}$$

【2014 BMVC】Speeding up Convolutional Neural Networks with Low Rank Expansions
【2015 ICLR】Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition
【2016 ICLR】Convolutional neural networks with low-rank regularization
【2016 ICLR】Training CNNs with Low-Rank Filters for Efficient Image Classification

Architecture

- **From 3D Convolution to Decoupled 1D and 2D**

$$v^{xyz} = \sum_m \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} \mathcal{K}^{pqr} v^{(x+p)(y+q)(z+r)}$$

Width * Height * Temporal but not cross channel

【2014 BMVC】Speeding up Convolutional Neural Networks with Low Rank Expansions
【2015 ICLR】Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition
【2016 ICLR】Convolutional neural networks with low-rank regularization
【2016 ICLR】Training CNNs with Low-Rank Filters for Efficient Image Classification

Architecture

- **From 3D Convolution to Decoupled 1D and 2D**

$$v^{xyz} = \sum_{m} \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} \mathcal{K}^{pqr} v^{(x+p)(y+q)(z+r)}$$

Width * Height * Temporal but not cross channel

- **Rewrite as** $\mathcal{O} = \mathcal{I} * \mathcal{K}$
- **Suppose we have** $\mathcal{K} = k_t \otimes K_{xy}$ **, Kronecker product**

【2014 BMVC】Speeding up Convolutional Neural Networks with Low Rank Expansions
【2015 ICLR】Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition
【2016 ICLR】Convolutional neural networks with low-rank regularization
【2016 ICLR】Training CNNs with Low-Rank Filters for Efficient Image Classification

Architecture

- **From 3D Convolution to Decoupled 1D and 2D**

$$v^{xyz} = \sum_{m} \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} \mathcal{K}^{pqr} v^{(x+p)(y+q)(z+r)}$$

Width * Height * Temporal but not cross channel

- **Rewrite as** $\mathcal{O} = \mathcal{I} * \mathcal{K}$
- **Suppose we have** $\mathcal{K} = k_t \otimes K_{xy}$ **, Kronecker product**

$\boxed{k_t \in \mathbb{R}^{n_t}}$ $\boxed{K_{xy} \in \mathbb{R}^{n_x \times n_y}}$

$$F_t(i_x, i_y, :) = \mathcal{I}(i_x, i_y, :) * k_t, \quad i_x = 1, 2, \cdots, m_x, \; i_y = 1, 2, \cdots, m_y.$$

$$F_{ts}(:, :, i_c) = F_t(:, :, i_c) * K_{xy}, \quad i_c = 1, 2, 3.$$

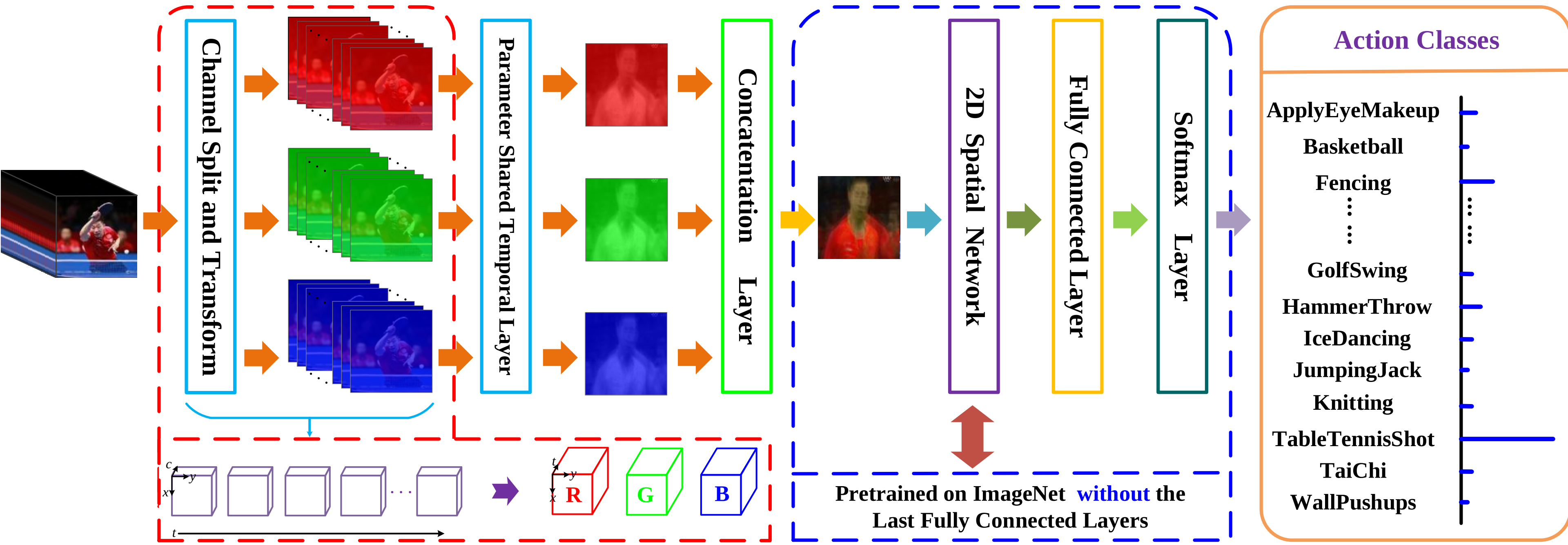
【2014 BMVC】Speeding up Convolutional Neural Networks with Low Rank Expansions
【2015 ICLR】Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition
【2016 ICLR】Convolutional neural networks with low-rank regularization
【2016 ICLR】Training CNNs with Low-Rank Filters for Efficient Image Classification

Cascaded Temporal Spatial Feature

$$\mathcal{O} = \mathcal{I} \circledast \mathcal{K}$$

$$= \sum_{c} \sum_{p=1}^{m_x} \sum_{q=1}^{m_y} \sum_{r=1}^{m_t} \mathcal{K}^{pqr} \mathcal{I}^{(i_x+p)(i_y+q)(i_z+r)}$$

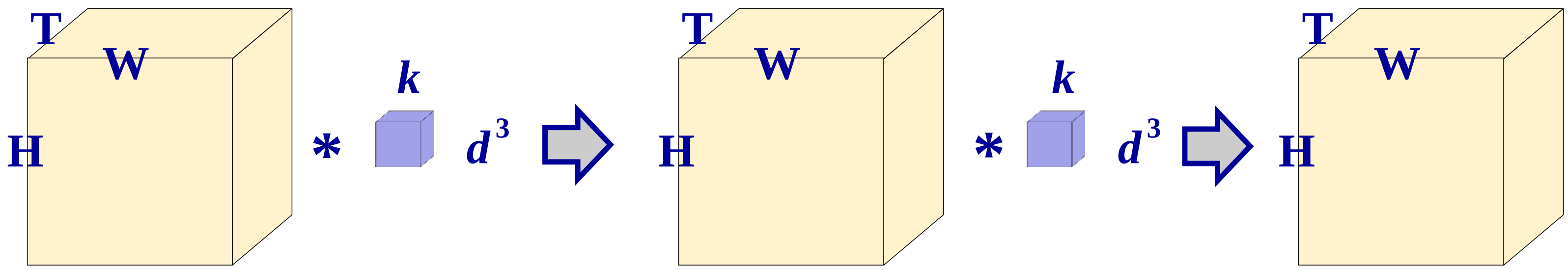
$$\mathcal{K} = k_t \otimes K_{xy}$$



$$F_t(i_x, i_y, :) = \mathcal{I}(i_x, i_y, :) * k_t, \quad i_x = 1, 2, \cdots, m_x, \quad i_y = 1, 2, \cdots, m_y.$$

$$F_{ts}(\,:\,,\,:\,, i_c) = F_t(\,:\,,\,:\,, i_c) * K_{xy}, \quad i_c = 1, 2, 3.$$

Cascaded Temporal Spatial Feature

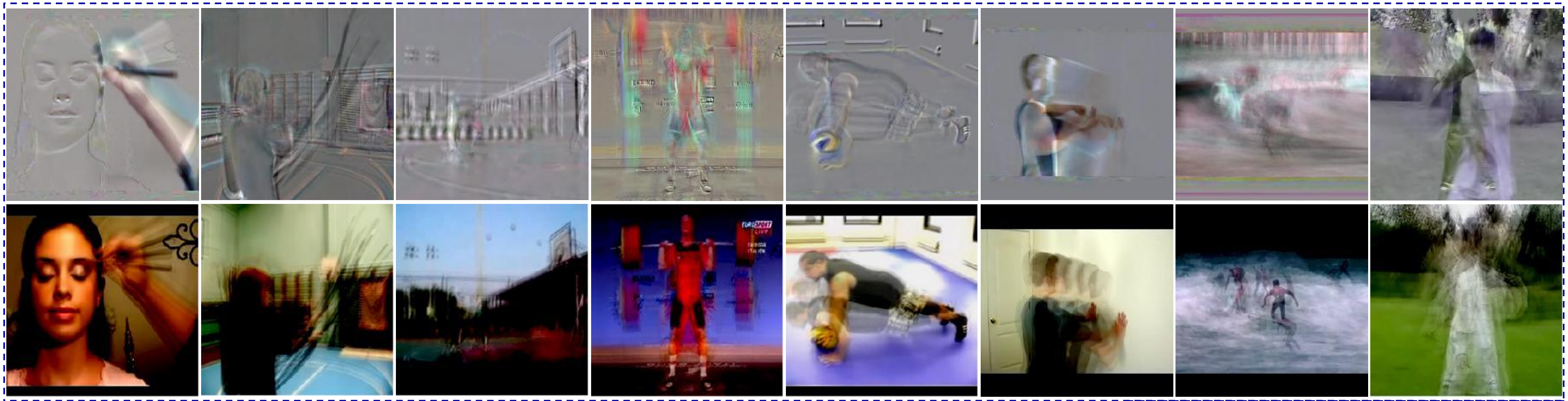


- Complexity Analysis

Criterion	3D	2D * 1D	1D * 2D
# Parameters	$2kd^3$	$kd(d+1)$	$kd(1+d)$
Computation	$k(1+k)WHTd^3$	$kd(d+k)WHT$	$kd(T+kd)WH$

Cascaded Temporal Spatial Feature

- Visualization of the Motion Image compared with Dynamic Image



- Complexity Analysis

Criterion	3D	2D * 1D	1D * 2D
# Parameters	$2kd^3$	$kd(d+1)$	$kd(1+d)$
Computation	$k(1+k)WHTd^3$	$kd(d+k)WHT$	$kd(T+kd)WH$

$$\mathbf{d}^* = \sum_t \alpha_t \psi(V_t)$$

$$F_t(i_x, i_y, :) = \sum_t \alpha_t \varphi(\mathcal{I}_t)$$

Cascaded Temporal Spatial Feature

➢ **Recognition Results： on UCF101 Action Dataset**

COMPARISON WITH STATE-OF-THE-ART METHODS ON UCF-101.

	Method	Accuracy
Trajectory Features	iDT [1]	0.762
	iDT + FV [2]	0.859
Pretrained CNN	ImageNet [12]	0.688
	CNN-M-2048 [9]	0.730
	VGG [13]	0.784
Single Image	Dynamic Image [3]	0.709
	Single Frame [4]	0.742
	Motion Image [5]	0.721
	Optical Flow [4]	0.823
CNN Features	ConvNet [14]	0.633
	LSTM [7]	0.758
	C3D [15]	0.815
	TSB-C3D [16]	0.827
	ResNet3D	0.826
	$F_{ST}CN$ [8]	0.845
	Two-Stream [9]	**0.869**
Fusion Image	Dynamic Image+Frame [3]	0.769
	Optical Flow+Frame [4]	0.859
	Motion Image+Frame [5]	**0.866**

Table 1. Comparison with Dynamic Image on UCF-101.

Method	Split1	Split2	Split3	Average
Mean Image	52.6%	53.4%	51.7%	52.6%
Max Image	48.0%	46.0%	42.3%	45.4%
Dynamic Image	57.2%	58.7%	57.7%	57.9%
Multi Dynamic Image	-	-	-	70.9%
Multi Dynamic Map	-	-	-	67.1%
Ours (without Aug)	44.9%	47.2%	43.7%	45.3%
Ours (with Aug)	**72.1%**	**72.6%**	**71.4%**	**72.1%**

➢ **Tricks:**

1、Data Augmentation.

2、Pre-Train on Sports-1M.

3、Video Level (vote).

4、Fusion with Frames.

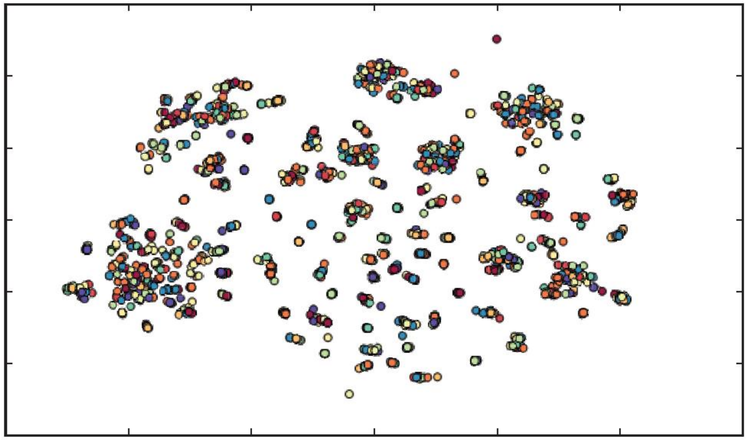
Cascaded Temporal Spatial Feature

➢ **Recognition Results： on HMDB51 Action Dataset**

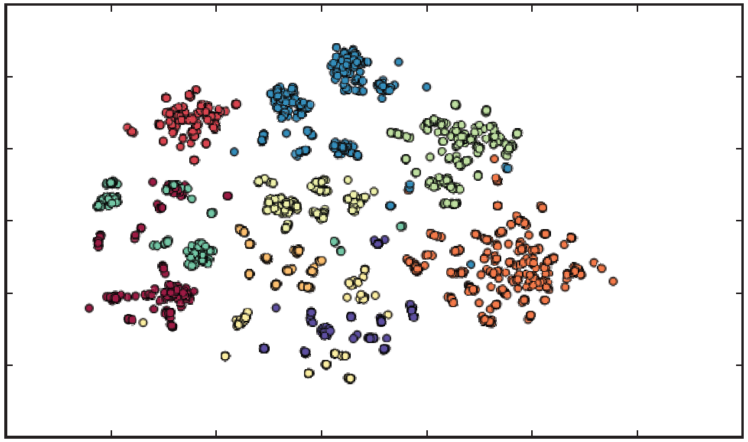
COMPARISON WITH STATE-OF-THE-ART METHODS ON HMDB-51.

	Method	Accuracy
Trajectory Features	iDT [1]	0.519
Single Image	Dynamic Image [3]	0.358
	Single Frame [4]	0.471
	Motion Image [5]	0.493
	Optical Flow [4]	**0.515**
CNN Features	VisualAttention [6]	0.413
	LSTM [7]	0.440
	ResNet3D	0.469
	F_{ST}CN [8]	0.490
Fusion Image	Dynamic Image+Frame [3]	0.428
	Two-Stream [9]	0.528
	Motion Image+Frame [5]	**0.529**

*Note that this table does not report the results combing CNN features with trajectory features.



(a) C3D



(b) Ours

Cascaded Temporal Spatial Feature

- **Recognition Results： on HMDB51 Action Dataset**

Truth: climb
C3D: climb
Ours: climb

Truth: catch
C3D: golf
Ours: catch

Truth: chew
C3D: chew
Ours: smoke

Truth: cartwheel
C3D: golf
Ours: golf

Truth: ride horse
C3D: ride horse
Ours: ride horse

Truth: smoke
C3D: laugh
Ours: smoke

Truth: drink
C3D: drink
Ours: eat

Truth: wave
C3D: sword
Ours: shot

Reference

[3] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould, “Dynamic image networks for action recognition,” in IEEE International Conference on Computer Vision and Pattern Recognition, 2016.

[4] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman,“Convolutional two-stream network fusion or video action recognition,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016.

[5] Tingzhao Yu, Huxiang Gu, Lingfeng Wang, Shiming Xiang, and Chunhong Pan, “Cascaded Temporal Spatial Features for Video Action Recognition,” in Proceedings of the IEEE Conference on Image Processing, 2017

[8] Karen Simonyan and Andrew Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[9] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi, “Human action recognition using factorized spatio-temporal convolutional networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4597–4605.

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in 2015 IEEE International Conference on Computer Vision . IEEE, 2015, pp. 4489–4497.

Thanks！