Paper ID 407

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

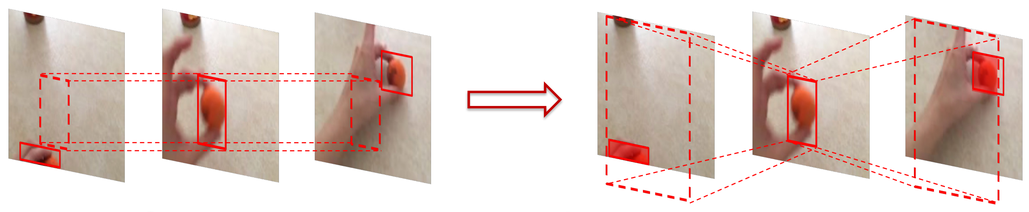
Hierarchical Hourglass Convolutional Network for Efficient Video Classification

Yi Tan, Yanbin Hao*, Hao Zhang, Shuo Wang, Xiangnan He*

• Motivation:

➢ Video dynamics result in **misalignment of visual clues** over temporal dimension

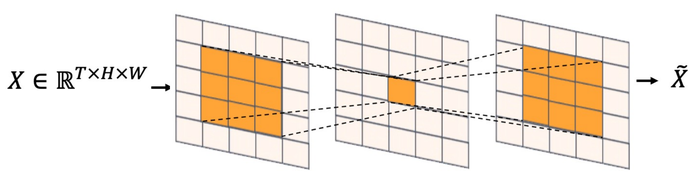
tackle visual displacements with hourglass shaped receptive field



Rigid temporal Conv may lose the target motion area

• Proposed framework:

➢ Hourglass Convolution (HgC): enlarging spatial receptive field for temporal neighbors



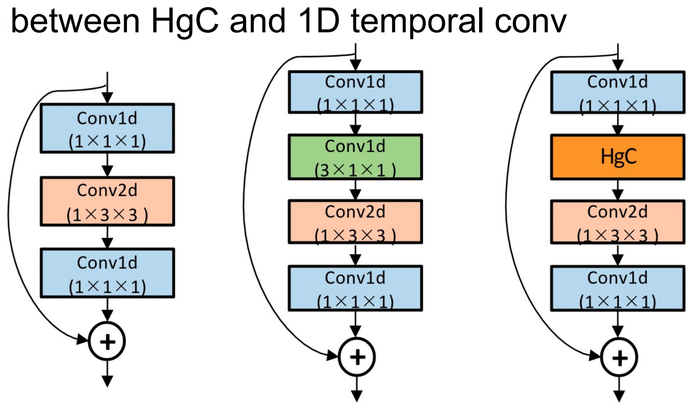
Hourglass convolution

$$HgC(\mathbf{X})_{t,h,w} = \sum_{i=-\frac{K}{2}}^{\frac{K}{2}} \alpha_i \cdot f(\mathbf{X}_{t+i,:,:}; W_{p\cdot|i|+1,p\cdot|i|+1})_{h,w},\ f(.) \text{ denotes spatial aggregation(e.g. conv \& pool),}$$

temporal offsets — temporal receptive field — Size of expanded spatial receptive field, p denotes the slope of expansion

✓ **By expanding spatial receptive field, HgC captures the spatial-temporal dynamics which vary their location, scale and pattern**

➢ Comparison between HgC and 1D temporal conv



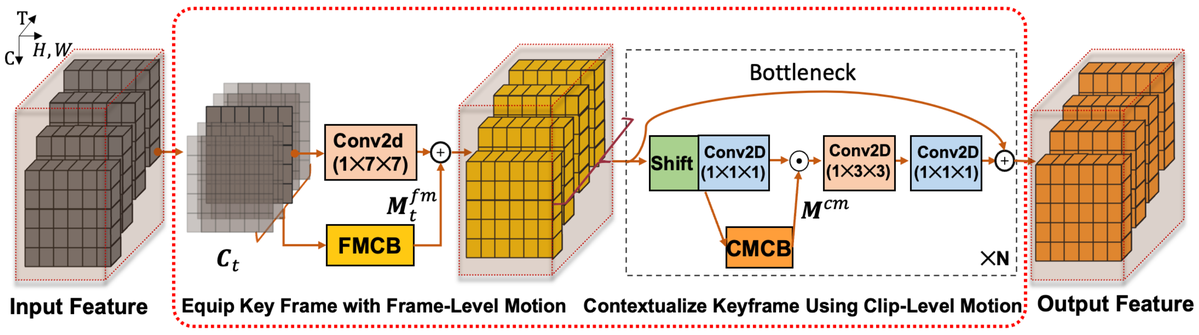
Method	$f(\bullet)$	Top-1	#P	FLOPs
TSN	——	19.7	23.9M	32.9G
R(2+1)D	——	46.0	23.9M	32.9G
HgC-ResNet	AvgPool2D	46.4	23.9M	32.9G
	Conv2D	**47.0**	23.9M	33.1G

➢ Architecture:

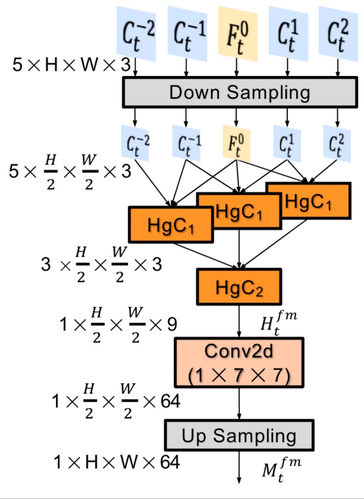
✓ Capture motion feature in different scales

a) Tiny motion between consecutive frames (**F**rame **M**otion **C**apture **B**lock)

b) Large movement between key frames (**C**lip **M**otion **C**apture **B**lock)

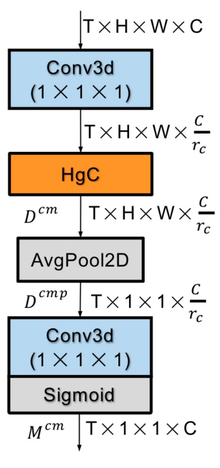


• FMCB and its effectiveness



method		top-1	top-5	#p	FLOPs
w/o FMCB		45.6	74.2	23.9M	32.9G
FMCB	p=2	52.3	80.3	23.9M	33.6G
	p=4	**52.5**	80.5	23.9M	33.6G
	p=6	52.3	80.3	23.9M	33.6G

• CMCB and its effectiveness



method		top-1	top-5	#p	FLOPs
w/o CMCB		52.5	80.5	23.9M	33.6G
CMCB	**p=2**	**53.6**	81.4	24.1M	33.8G
	p=4	53.4	81.4	24.1M	33.8G
	p=6	**53.6**	81.8	24.1M	33.8G

• Comparison with SOTA:

✓ Something-Something V1&V2

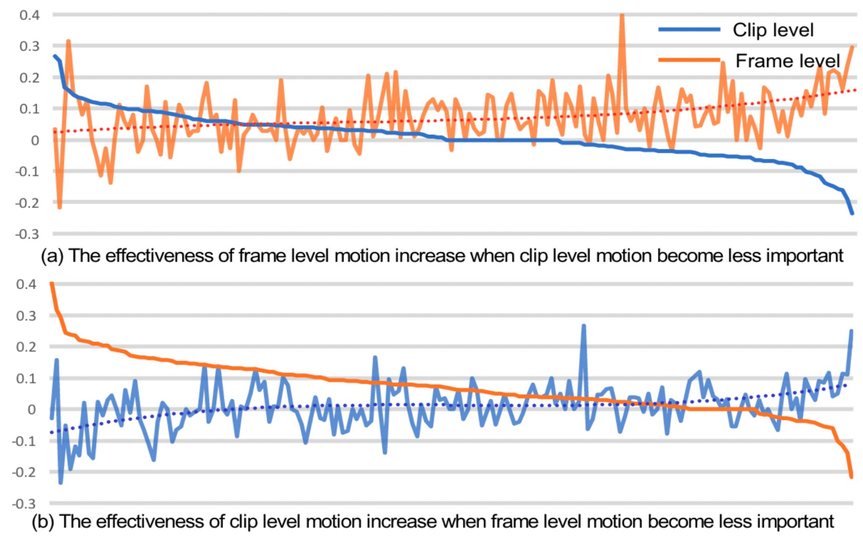
Method	Backbone	Keyframes×Views	FLOPs	V1		V2	
				Top-1	Top-5	Top-1	Top-5
I3D [3]	3DResNet-50	32×2	153.0G×2	41.6	72.2	—	—
NLI3D [51]			168.0G×2	44.4	76	—	—
NLI3D+GCN [52]			303.0G×2	46.1	76.8	—	—
GST [34]	ResNet-50	16×1	59.0G×1	48.6	77.9	62.6	87.9
TSM [30]	ResNet-50	16×1×2	65.8G×1×2	48.4	78.1	63.1	88.2
SDA-TSM [44]		16×1×2	67.8G×1×2	52.2	80.9	64.7	89.5
TIN [39]	ResNet-50	16×1	67.0G×1	47	76.5	60.1	86.4
TEINet [31]	ResNet-50	16×1	66.0G×1	49.9	—	62.1	—
TAM [33]	ResNet-50	16×1	66.0G×1	47.6	77.7	62.5	87.6
TEA [27]	ResNet-50	16×30	70.0G×30	52.3	81.9	—	—
STM [22]	ResNet-50	8×30	33.3G×30	49.2	79.3	62.3	88.8
STM [22]		16×30	66.5G×30	50.7	80.4	64.2	89.8
MoViNet-A3 [24]	—	50	23.7G	—	—	64.1	88.8
TDN [49]	ResNet-50	(8+16)×1	108.0G×1	55.1	82.9	67.0	89.5
SELFYNet [25]	ResNet-50	8×1	37.0G×1	52.5	80.8	64.5	89.4
SELFYNet [25]		16×1	77.0G×1	54.3	82.9	65.7	89.8
SELFYNet [25]		(8+16)×1	114.0G×1	55.8	83.9	67.4	91.0
TimeSformer-HR [2]	Transformer	16×3	1703G×3	—	—	62.5	—
ViViT-L [1]		32×4	903G×4	—	—	65.4	89.8
MViT-B [8]		64×3	455G×3	—	—	67.7	90.9
Video-Swin-B [32]		16×3	321G×3	—	—	**69.6**	**92.7**
H²CN(ours)	ResNet-50	8×1	33.8G×1	53.6	81.4	65.2	89.7
H²CN(ours)		16×1	67.6G×1	55.0	82.4	66.4	90.1
H²CN(ours)		(8+16)×1	101.4G×1	**56.7**	**83.2**	**67.9**	**91.2**

✓ Kinetics-400

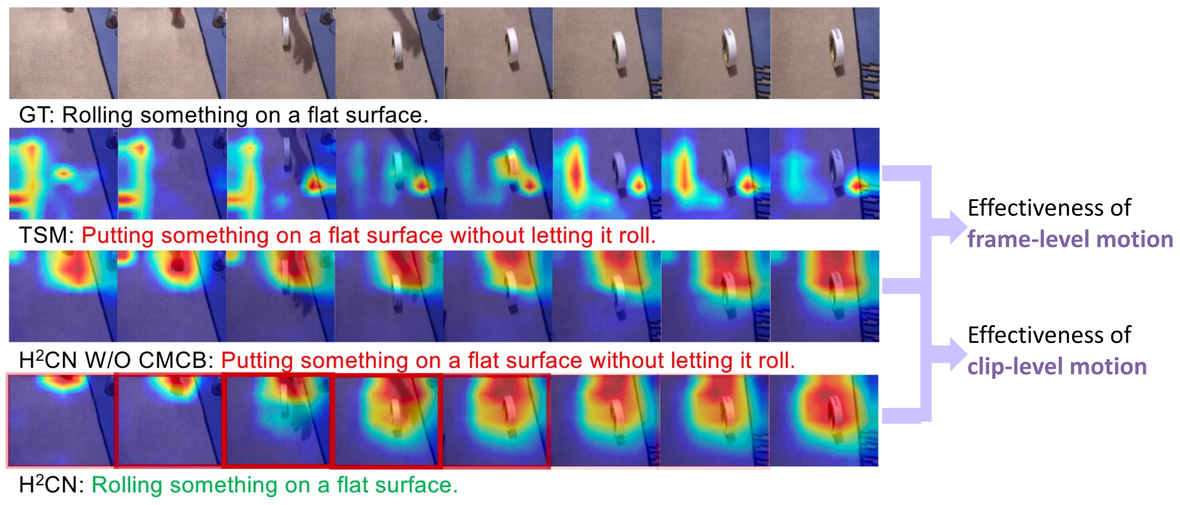
Method	Backbone	Frames	GFLOPs	Top1	Top5
TSN [50]	InceptionV3	25	80×10	72.5	90.2
TSM [30]	ResNet50	16	65×30	74.7	91.4
I3D [3]	InceptionV1	64	—	72.1	90.3
R(2+1)D [47]	ResNet34	32	152×10	74.3	91.4
S3D-G [54]	InceptionV1	64×30	71.4×30	74.7	93.4
NL-I3D [51]	ResNet50	32	282×10	74.9	91.6
TEA [27]	ResNet50	16	70×30	76.1	92.5
TANet [33]	ResNet50	16	86×12	76.9	92.9
SmallBigNet [26]	ResNet50	8	57×30	76.3	92.5
SlowFast [12]	ResNet50	8+32	65.7×30	77.0	92.6
X3D-L [11]	—	16	24.8×30	77.5	92.9
MoViNet-A5 [24]	—	120	289	78.2	—
SELFYNet [25]	ResNet50	16	77×30	77.1	—
TDN [49]	ResNet50	8+16	108×30	78.4	93.6
H²CN (Ours)	ResNet50	8	33.8×30	76.9	93.0
H²CN (Ours)	ResNet50	16	67.6×30	77.9	93.3
H²CN (Ours)	ResNet50	8+16	101.4×30	**78.7**	**93.6**

• Visualization

➢ Motion information on different levels works in a **complementary** way



➢ Spatiotemporal response of TSM (Backbone), H²CN w/o CMCB, and H²CN



• Contact & Resources

ty133@mail.ustc.edu.cn, haoyanbin@hotmail.com
{zhanghaoinf, shuowang.hfut, xiangnanhe}@gmail.com

GitHub