


Compressed Video Action Recognition

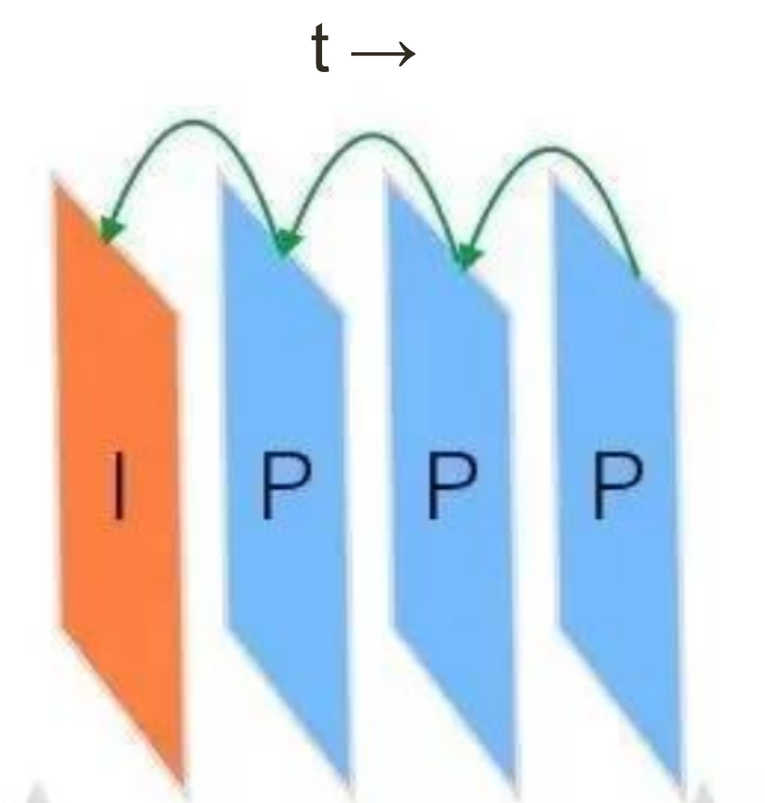
Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, Philipp Krähenbühl

CVPR18 Oral Spotlight, code

未来媒体研究中心 CENTER FOR FUTURE MEDIA
电子科技大学 University of Electronic Science and Technology of China

Main idea

- Use compressed video (MPEG4-coded) as network input.
  - ✔ Less redundant
  - ✔ More temporal



MPEG4 coding

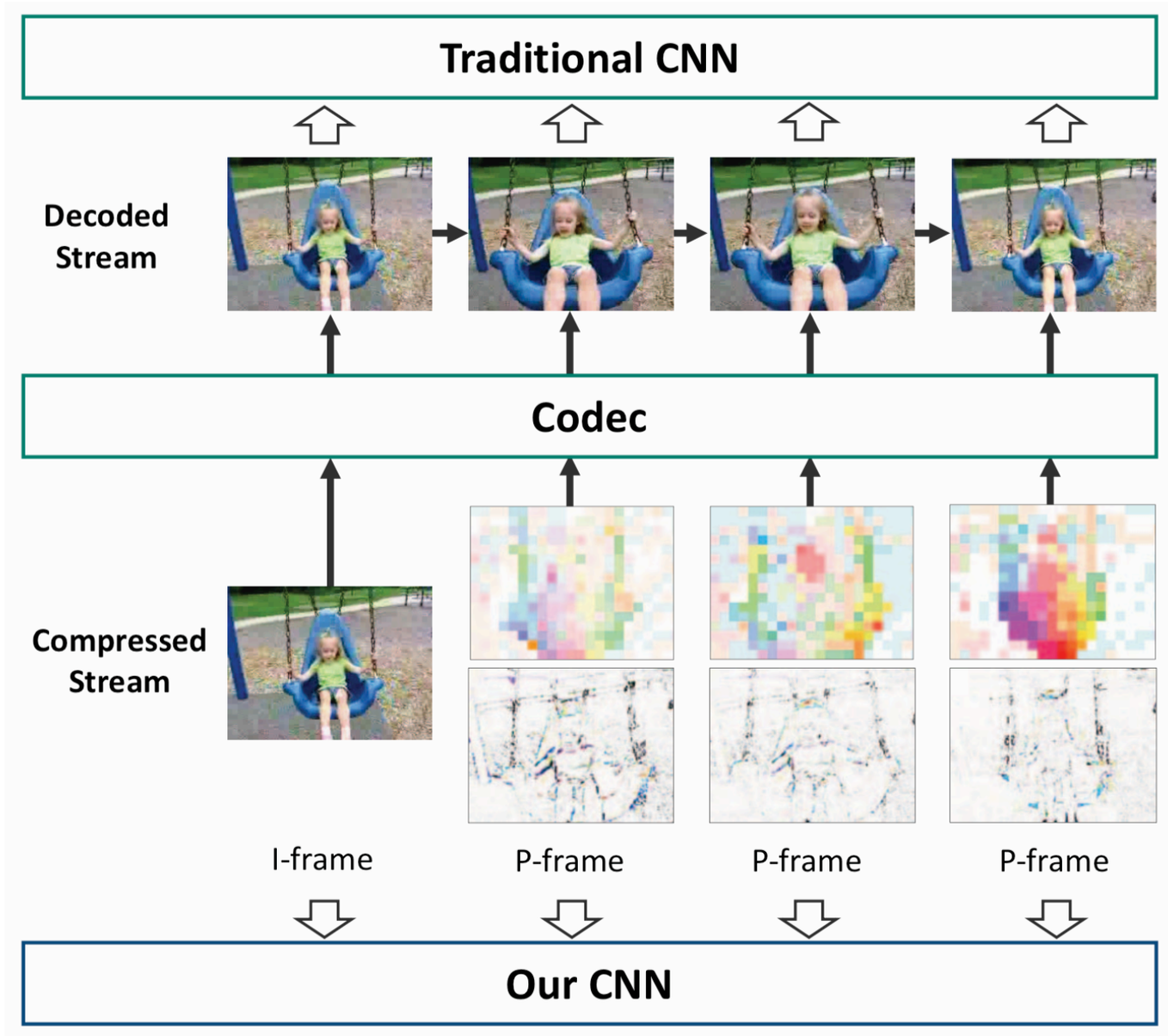
I : RGB Image

P : Motion vector + Residual

$$I_i^{(t)} = I_{i-\mathcal{T}_i^{(t)}}^{(t-1)} + \Delta_i^{(t)},$$

Main idea

- Traditional inputs:
  - RGB frames
- Proposed inputs:
  - I-frame,
  - Motion vector
  - Residual

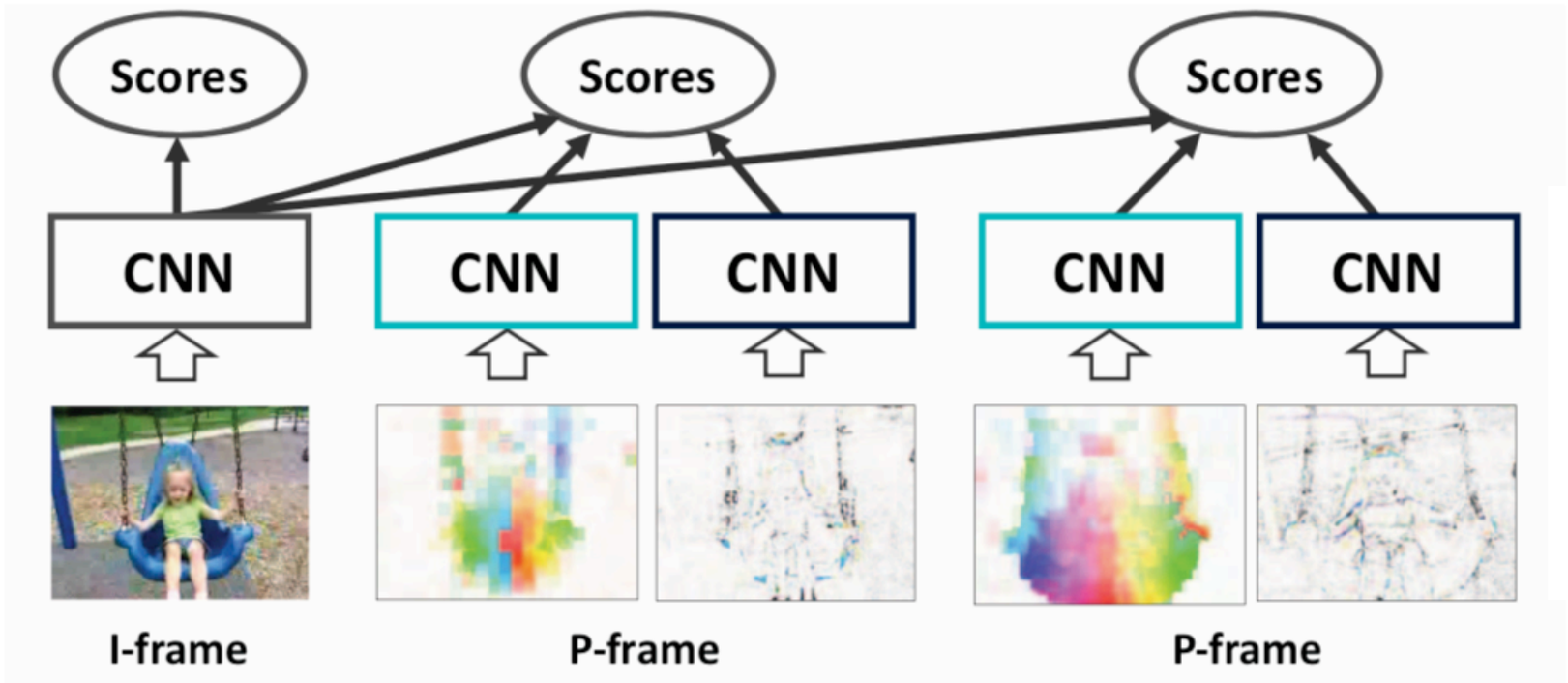


Claim

- Compressed video
  - have higher information density
  - have nicer temporal structural
  - doesn’t need to be decompressed
  - lead to faster, simpler and more accurate model.

Model

- ResNet152 for I-frame and ResNet18 for P-frame
- Add fusion (summing of scores works best)



$$x^{(0)}_{\mathrm{RGB}} := \phi_{\mathrm{RGB}}\left(I^{(0)}\right)$$
$$x^{(t)}_{\mathrm{motion}} := \phi_{\mathrm{motion}}\left(\mathcal{D}^{(t)}\right)$$
$$x^{(t)}_{\mathrm{residual}} := \phi_{\mathrm{residual}}\left(\mathcal{R}^{(t)}\right)$$

Figure 5: Decoupled model. All networks can be trained independently. Models are shared across P-frames.

Accumulated MV & Residual

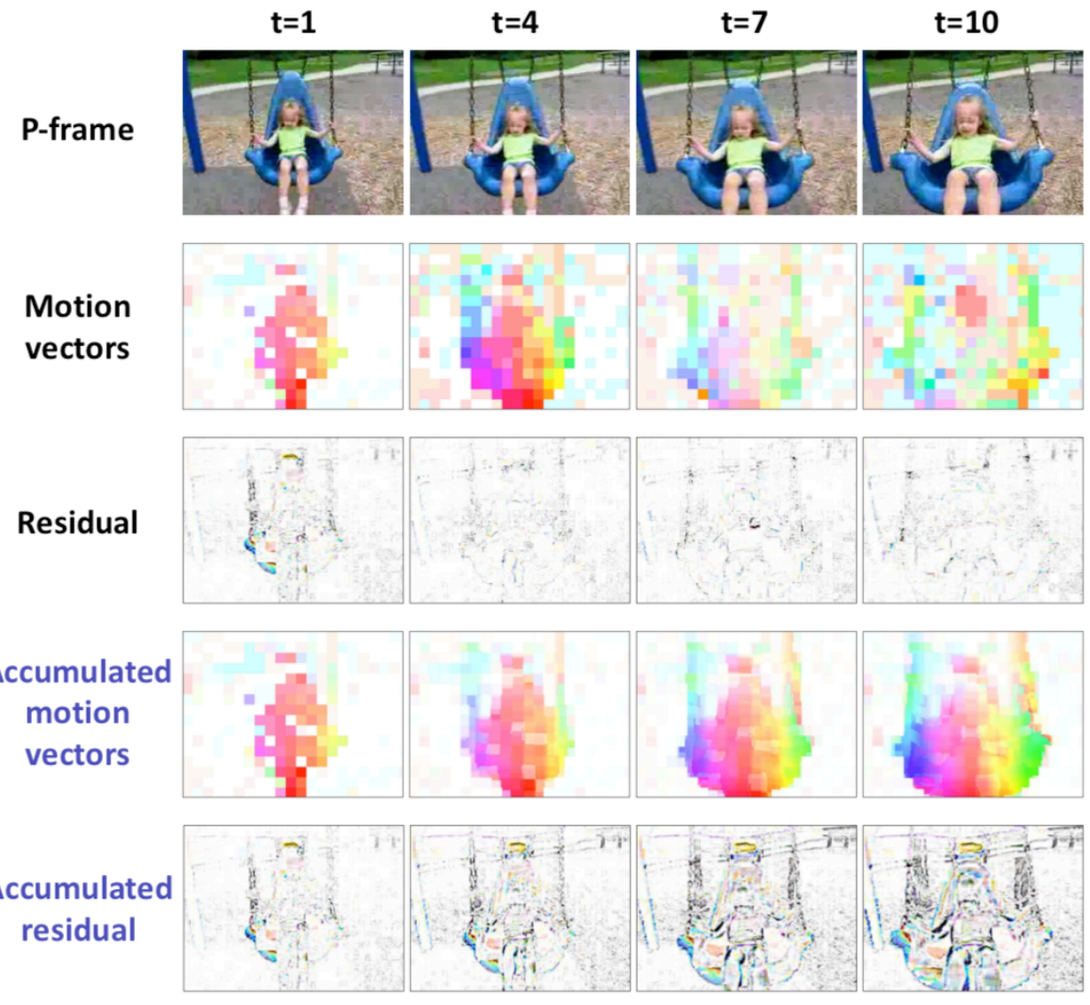


Figure 2: Original motion vectors and residuals describe only the change between two frames. Usually the signal to noise ratio is very low and hard to model. The accumulated motion vectors and residuals consider longer term difference and show clearer patterns. Assume I-frame is at $t = 0$. Motion vectors are plotted in HSV space, where the

- Motivation: CNN friendly
  - Acc. MV resemble Optic Flow
  - Acc. Residual resemble Image

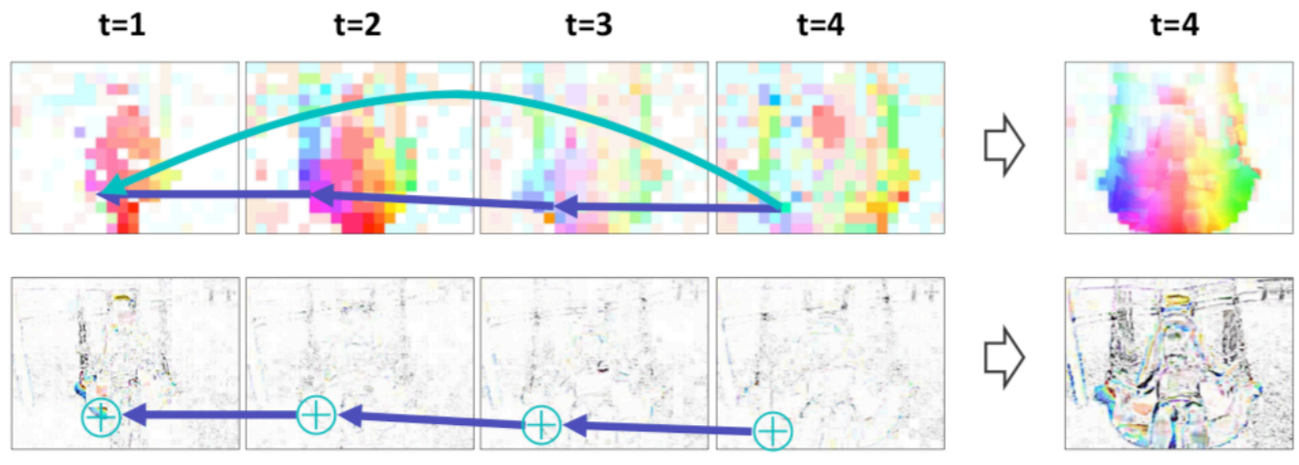
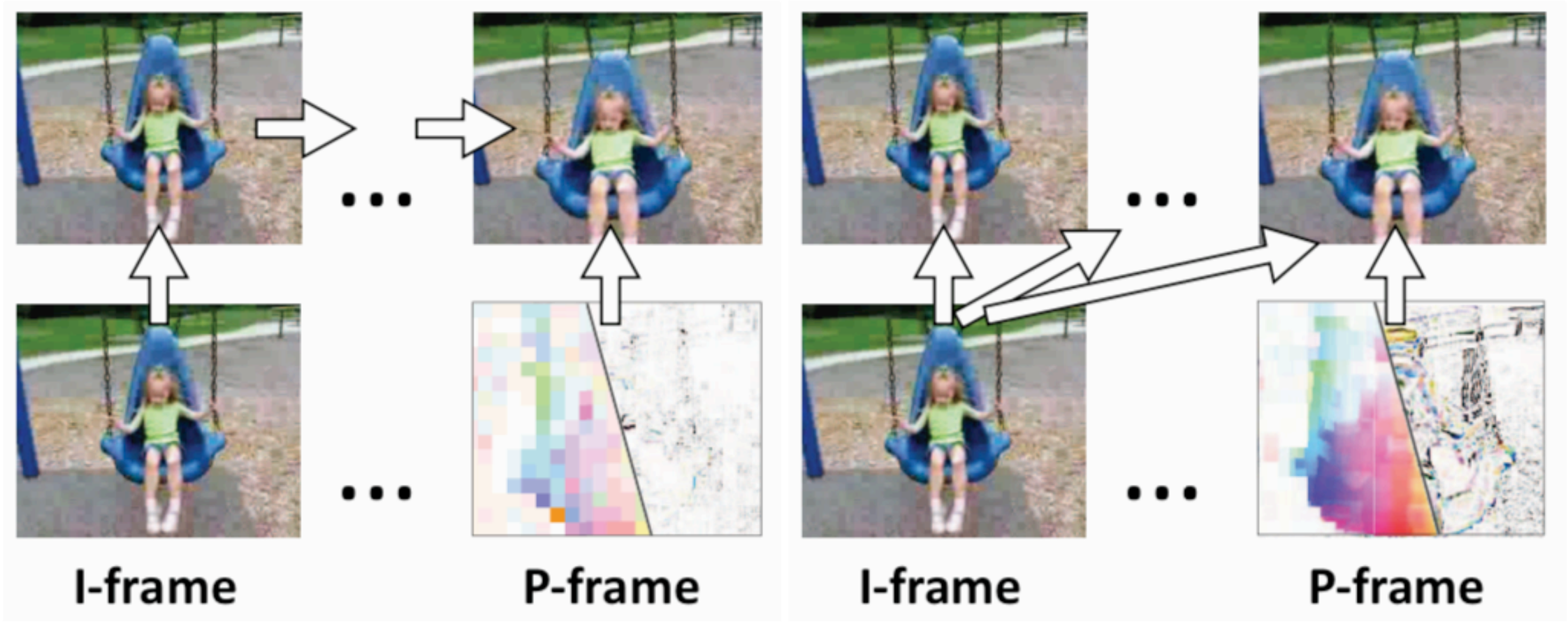


Figure 3: We trace all motion vectors back to the reference I-frame and accumulate the residual. Now each P-frame depends only on the I-frame but not other P-frames.

Decouple frames (parallel)



(a) Original dependency. (b) New dependency.

Figure 4: We decouple the dependencies between P-frames so that they can be processed in parallel.

Effectiveness of M & R

	I	M	R	I+M	I+R	I+M+R (gain)
UCF-101						
Split 1	88.4	63.9	79.9	<u>90.4</u>	90.0	**90.8** (+2.4)
Split 2	87.4	64.6	80.8	<u>89.9</u>	89.6	**90.5** (+3.1)
Split 3	87.3	66.6	82.1	<u>89.6</u>	89.4	**90.0** (+2.7)
Average	87.7	65.0	80.9	<u>89.9</u>	89.7	**90.4** (+2.7)
HMDB-51						
Split 1	54.1	37.8	44.6	<u>60.3</u>	55.9	**60.4** (+6.3)
Split 2	51.9	38.7	43.1	<u>57.9</u>	54.2	**58.2** (+6.3)
Split 3	54.1	39.7	44.4	<u>58.5</u>	55.6	**58.7** (+4.6)
Average	53.3	38.8	44.1	<u>58.9</u>	55.2	**59.1** (+5.8)

Table 1: Action recognition accuracy on UFC-101 [34] and HMDB-51 [18]. Here we compare training with different sources of information. "+" denotes score fusion of models. I: I-frame RGB image. M: motion vectors. R: residuals. The bold numbers indicate the best and the underlined numbers indicate the second best performance.

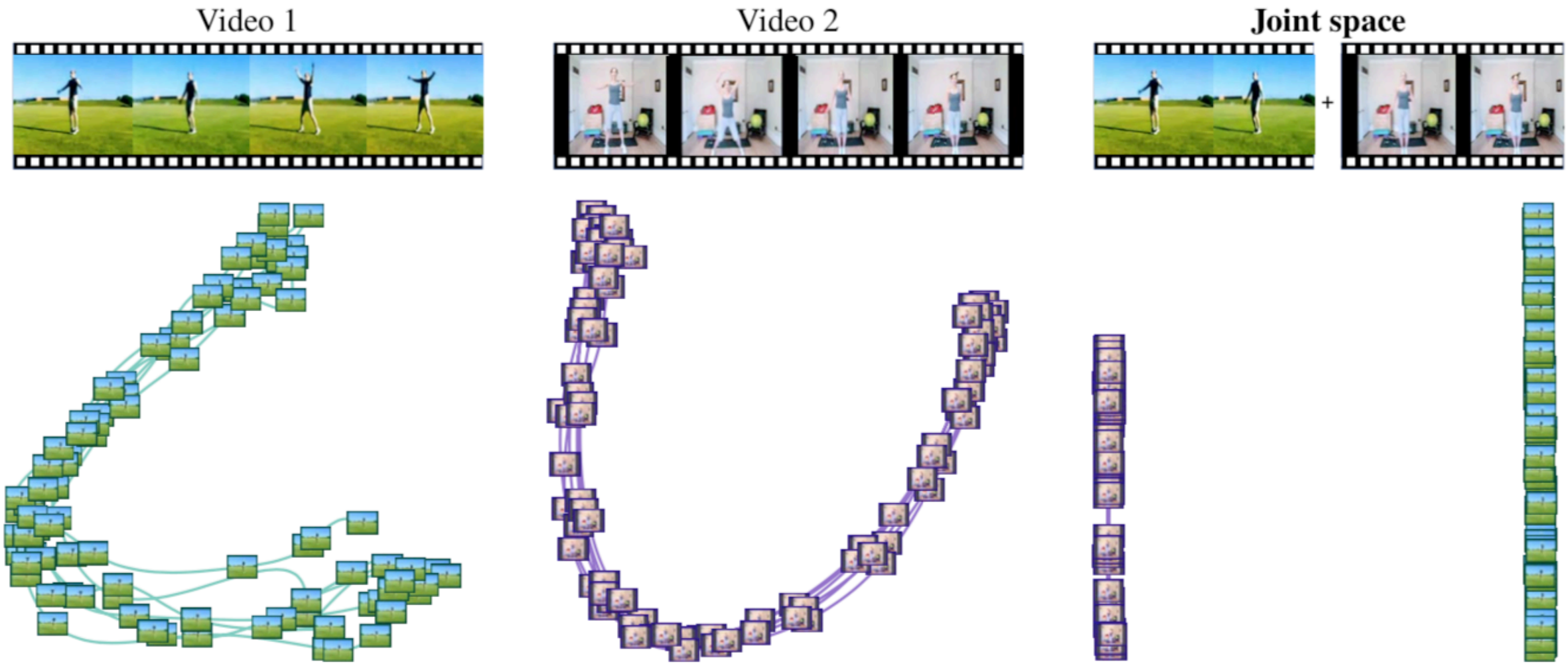
Accumulation result

	M	R	I+M	I+R	I+M+R
Original	58.3	79.0	90.0	89.8	90.4
Accumulated	**63.9**	**79.9**	**90.4**	**90.0**	**90.8**

Table 2: Action recognition accuracy on UFC-101 [34] (Split 1). The two rows show the performance of the models trained using the original motion vectors/residuals and the models using the accumulated ones respectively. I: I-frame RGB image. M: motion vectors. R: residuals.

t-SNE of ‘Jumping Jack’ in RGB

- ✗ Intra-class separation
- ✗ None-repeated patterns



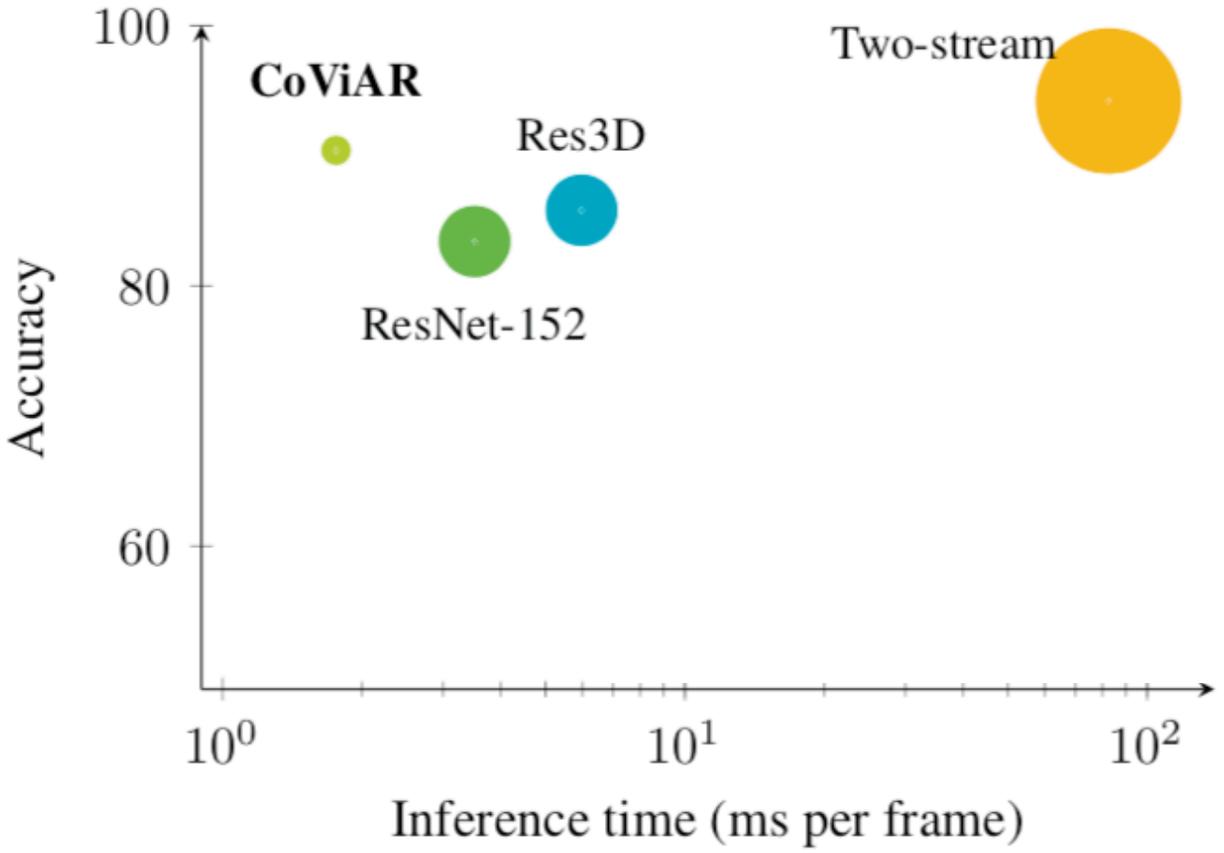
t-SNE of ‘Jumping Jack’ in MV

- ✓ Intra-class congregation
- ✓ Circular pattern

t-SNE of ‘Jumping Jack’ in R

- ✓ Intra-class congregation
- ✓ Circular pattern

Speed, Accuracy and Data size



	GFLOPs	Accuracy (%) UCF-101	HMDB-51
ResNet-50 [8]	**3.8**	82.3	48.9
ResNet-152 [8]	11.3	83.4	46.7
C3D [39]	38.5	82.3	51.6
Res3D [40]	19.3	85.8	54.9
CoViAR	4.2	**90.4**	**59.1**

Table 3: Network computation complexity and accuracy of each method. Our method is 4.6x more efficient than state-of-the-art 3D CNN, while being much more accurate.

	Preprocess	CNN (sequential)	CNN (concurrent)
Two-stream			
BN-Inception	75.0	1.6	0.9
ResNet-152	75.0	7.5	4.0
CoViAR	**2.87/0.46**	**1.3**	**0.3**

Table 4: Speed (ms) per frame. CoViAR is fast in both preprocessing and CNN computation. Its preprocessing speed is presented for both single-thread / multi-thread settings.

Compare with STOA

	UCF-101	HMDB-51
Without optical flow		
Karpathy *et al.* [16]	65.4	-
ResNet-50 [12] (from ST-Mult [8])	82.3	48.9
ResNet-152 [12] (from ST-Mult [8])	83.4	46.7
C3D [39]	82.3	51.6
Res3D [40]	85.8	54.9
TSN (RGB-only) [44]*	85.7	-
TLE (RGB-only) [5]†	87.9	54.2
I3D (RGB-only) [2]*	84.5	49.8
MV-CNN [49]	86.4	-
P3D ResNet [27]	88.6	-
Attentional Pooling [10]	-	52.2
CoViAR	**90.4**	**59.1**
With optical flow		
iDT+FV [42]	-	57.2
Two-Stream [33]	88.0	59.4
Two-Stream fusion [9]	92.5	65.4
LRCN [6]	82.7	
Composite LSTM Model [35]	84.3	44.0
ActionVLAD [11]	92.7	66.9
ST-ResNet [7]	93.4	66.4
ST-Mult [8]	94.2	68.9
I3D [2]*	93.4	66.4
TLE [5]†	93.8	68.8
L^2STM [37]	93.6	66.2
ShuttleNet [30]	94.4	66.6
STPN [45]	**94.6**	68.9
TSN [44]	94.2	69.4
CoViAR + optical flow	**94.9**	**70.2**

Table 6 (ImageNet pretrained)

		mAP (%)	wAP (%)
Without optical flow			
ActionVLAD [11] (RGB only)		17.6	25.1
Sigurdsson *et al.* [31] (RGB only)	extra supervision	18.3	-
CoViAR		**21.9**	**29.4**
With optical flow			
Two-stream [33] (from [32])		14.3	-
Two-stream [33] + iDT [42] (from [32])		18.6	-
ActionVLAD [11] (RGB only) + iDT		21.0	29.9
Sigurdsson *et al.* [31]		22.4	-
CoViAR + optical flow		**24.1**	**32.3**

Table 7: Accuracy on Charades [32]. Without using additional annotations as Sigurdsson *et al.* [31], our method achieves the best performance.

Incorporating with Optic Flow

	UCF-101			HMDB-51		
	CoViAR	Flow	CoViAR +flow	CoViAR	Flow	CoViAR +flow
Split 1	90.8	87.7	**94.0**	60.4	61.8	**71.5**
Split 2	90.5	90.2	**95.4**	58.2	63.7	**69.4**
Split 3	90.0	89.1	**95.2**	58.7	64.2	**69.7**
Average	90.4	89.0	**94.9**	59.1	63.2	**70.2**

Table 5: Action recognition accuracy on UFC-101 [34] and HMDB-51 [18]. Combining our model with a temporal-stream network achieves state-of-the-art performance.

Thank You!

未来媒体研究中心
CENTER FOR FUTURE MEDIA

电子科技大学
University of Electronic Science and Technology of China