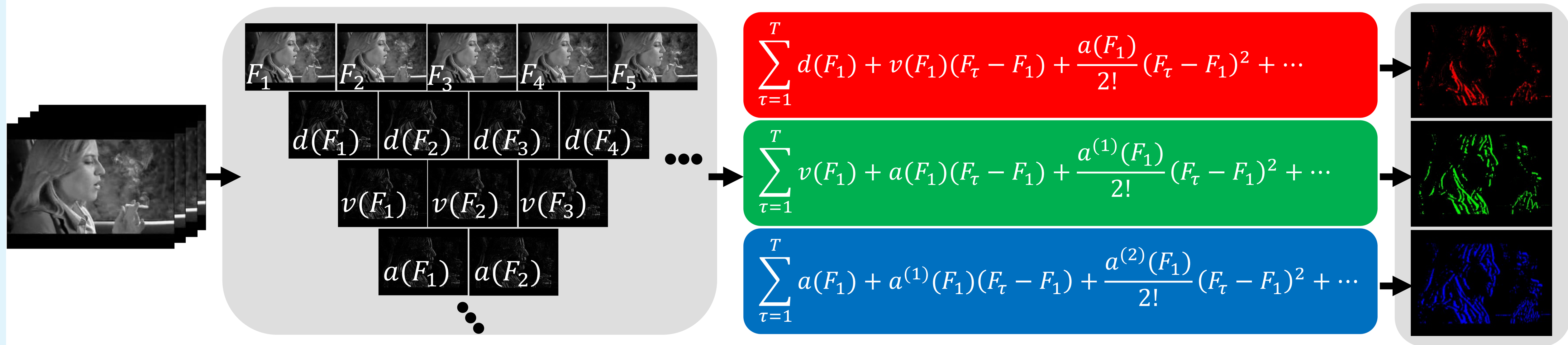
Taylor Videos for Action Recognition

Lei Wang[1] Xiuyuan Yuan[1] Tom Gedeon[2] Liang Zheng[1]

[1]Australian National University [2]Curtin University

Motivation and key ideas

Taylor series locally approximates non-linear functions: $f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$.



Our motion extraction function: $f(\boldsymbol{F}_T) = \sum_{k=0}^{\infty} \frac{f^{(k)}(\boldsymbol{F}_1)}{k!} \odot (\boldsymbol{F}_T - \boldsymbol{F}_1)^{\circ k}$.

Combining short-term and long-term motions in a temporal block: $\boldsymbol{M}_f = \frac{1}{T}\sum_{\tau=1}^{T} f(\boldsymbol{F}_\tau)$.

Subscript f is used to denote extracting a certain motion concept: displacement, velocity, and acceleration.

Quantitative results

	Model	Pretrain	Input	HMDB-51	CATER static	CATER moving	MPII
2D CNNs	TSN	ImageNet	RGB	54.9	49.6	51.6	38.4
			Taylor	56.4	73.8	62.7	42.2
	TSM	ImageNet	RGB	-	79.9	65.8	46.7
			GrayST	-	82.2	74.7	48.7
			Taylor	-	83.1	75.5	50.1
3D CNNs	I3D	ImageNet	RGB	49.8	73.5	57.7	42.8
			Taylor	65.2	74.7	60.5	43.0
		Kinetics	RGB	74.3	75.4	61.9	48.7
			OPT	77.3	78.5	66.3	51.0
			Taylor	78.1	80.2	69.8	52.3
	R(2+1)D	Sports1M	RGB	66.6	-	-	-
			Taylor	67.4	-	-	-
Transf.	TimeSformer	Kinetics	RGB	71.7	69.9	57.6	41.0
			Taylor	72.1	71.2	58.2	42.8
	Swin Transformer	Kinetics	RGB	72.9	72.2	63.5	46.6
			Taylor	73.5	73.0	64.7	47.0

Table 1: Comparing the Taylor video with other input modalities on three datasets with various action recognition models and pre-training datesets.

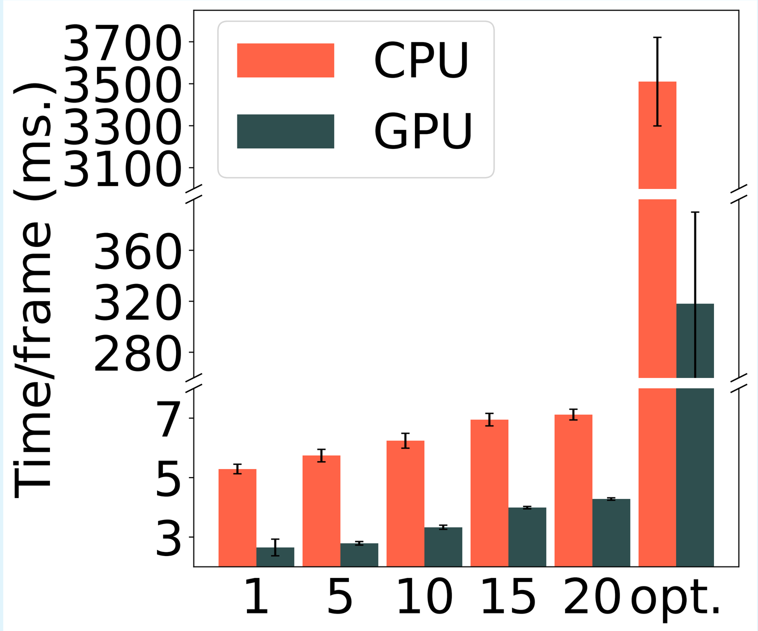


Figure 1: Time cost (milliseconds, ms.) of computing a single Taylor frame (with 1, 5, 10, 15, 20 terms) and optical flow (opt.)

Model	Input	K400	K600	SSv2
TSM	RGB	76.3	-	63.4
	Taylor	77.6	-	65.1
I3D	RGB	77.7	-	-
	Taylor	79.3	-	-
TimeSformer	RGB	80.7	82.2	62.5
	Taylor	81.5	83.1	63.7
VideoMAE	RGB	79.8	-	69.3
	Taylor	80.4	-	70.0
Swin Transformer	RGB	-	-	69.6
	Taylor	-	-	71.1

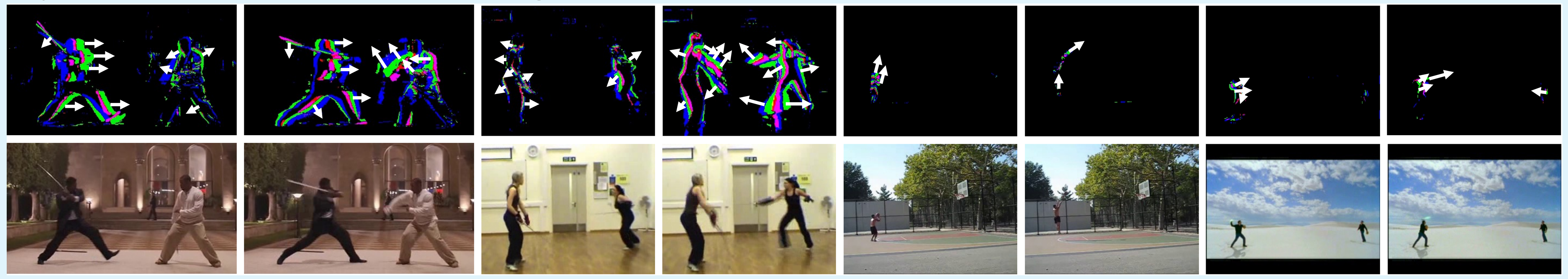
Table 2: Evaluations of Taylor videos on large-scale Kinetics (K400 / K600) and Something-Something v2 (SSv2).

Model	Input	NTU-60 X-Sub	NTU-60 X-View	NTU-120 X-Sub	NTU-120 X-Set	K-Skel Top-1
ST-GCN	Skeleton	81.5	88.3	70.7	73.2	30.7
	Taylor	85.4	93.0	78.5	80.1	35.1
InfoGCN	Skeleton	93.0	97.1	89.8	91.2	-
	Taylor	94.6	98.5	91.6	93.7	-
AGE-Ens	Skeleton	91.0	96.1	87.6	88.8	-
	Taylor	95.0	98.3	91.8	92.5	-
3Mformer	Skeleton	94.8	98.7	92.0	93.8	48.3
	Taylor	95.3	98.8	92.6	94.7	49.2

Table 3: Comparing Taylor-transformed skeletons with original skeletons on NTU-60, NTU-120 and Kinetics-Skeleton (K-Skel).

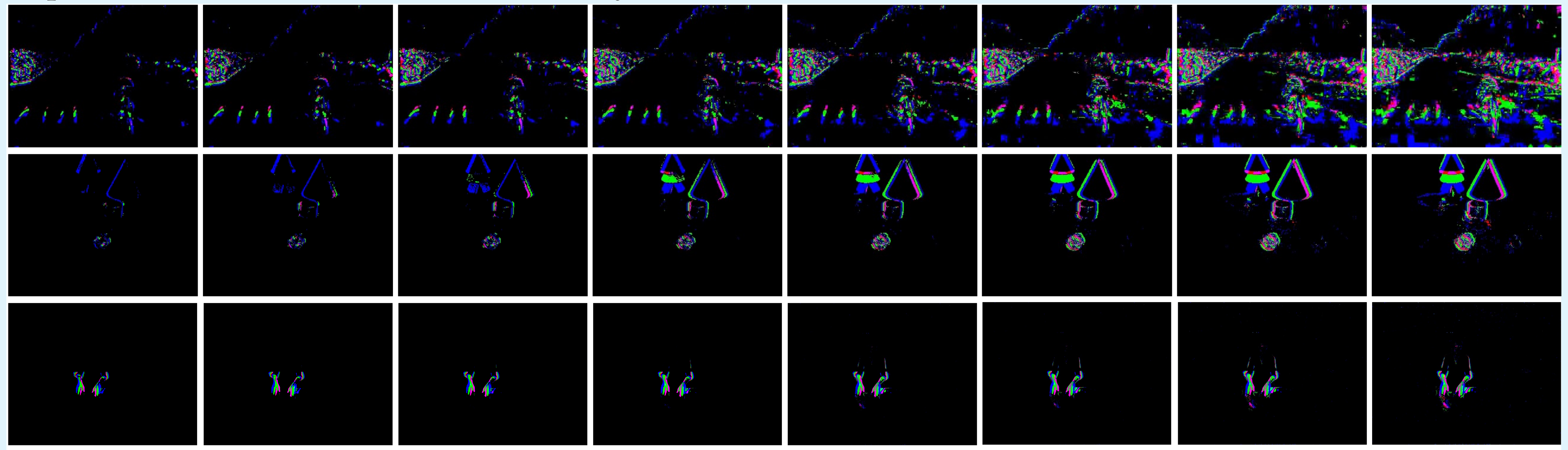
Qualitative results

Taylor frames indicate motion strengths and directions.



Taylor videos remove redundancy, such as static backgrounds, unstable pixels, watermarks, and captions.

Impact of the number of terms used in Taylor series.



1 term | 2 terms | 3 terms | 5 terms | 7 terms | 10 terms | 15 terms | 20 terms

Taylor videos are able to remove distinct facial features of individuals compared to RGB videos. This allows the data collection and processing to have improved privacy.