Enhancing Video Understanding with New Representation & Fusion

Lei Wang

Griffith University

February 25, 2025

Table of Contents

Action Recognition, Challenges & Benchmarks

Action Recognition, Challenges & Benchmarks

Action Recognition: recognize/identify actions in video

Motivations:

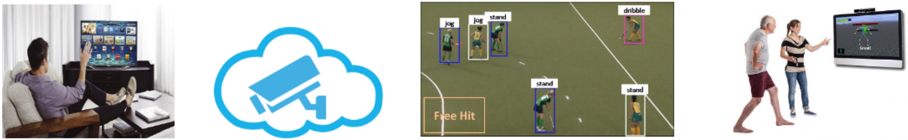


Figure 1: Many useful applications.

Challenges:

Figure 2: Many challenging issues.

Action Recognition, Challenges & Benchmarks (cont.)

Table 1: Some benchmarks for action recognition.

Datasets	Year	Classes	Subjects	#views	#video clips	Sensor	Modalities
MSRAction3D	2010	20	10	1	567	Kinect v1	Depth+3DJoints
3D Action Pairs	2013	12	10	1	360	Kinect v1	RGB+Depth+3DJoints
UWA3D Activity	2014	30	10	1	701	Kinect v1	RGB+Depth+3DJoints
UWA3D Multiview Activity II	2015	30	9	4	1,070	Kinect v1	RGB+Depth+3DJoints
MPII Cooking Activities	2012	64	12	1	3,748	-	RGB
HMDB-51	2011	51	-	-	6,766	-	RGB
EPIC-Kitchens	2018	149	32	-	39,594	-	RGB+Flow
NTU RGB+D	2016	60	40	80	56,880	Kinect v2	RGB+Depth+IR+3DJoints
Charades	2016	157	-	-	66,500	-	RGB+Flow
NTU RGB+D 120	2019	120	106	155	114,480	Kinect v2	RGB+Depth+IR+3DJoints
Kinetics-skeleton	2017	400	-	-	260,232	-	2DJoints
Kinetics	2018	400	-	-	$\sim$ 300,000	-	RGB

Figure 3: Video frame images

[A] A comparative review of recent kinect-based action recognition algorithms. TIP'20.

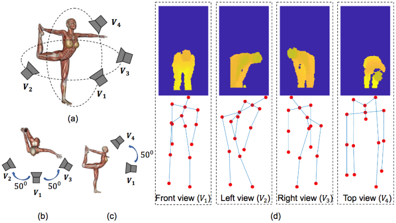


Figure 4: Setup, depth frames & skeletons[A].

Action Recognition, Challenges & Benchmarks (cont.)

(a) cut apart (b) cut dice (c) cut slices (d) cut slices

Figure 5: Finegrained action recognition (MPII Cooking Activities)

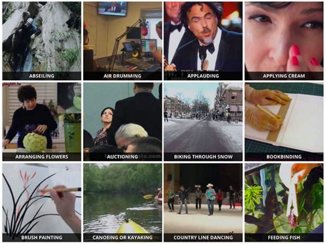


Figure 6: Video frames from Kinetics700

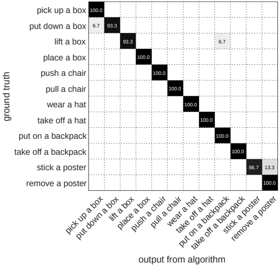


Figure 7: Confusion matrix

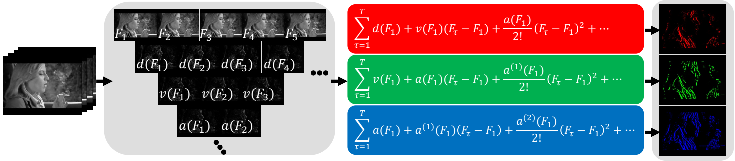
A New Video Representation: Taylor Videos

Motivation and key ideas

- Taylor series locally approximates non-linear functions. It is an infinite sum of terms expressed in terms of the function's derivatives at a single point:

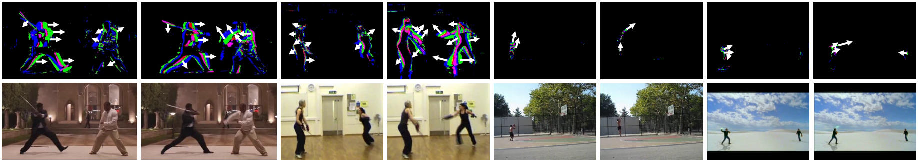
$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k. \tag{1}$$

- The first few terms of the series can reconstruct most of $f(x)$.
- 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.



Qualitative results

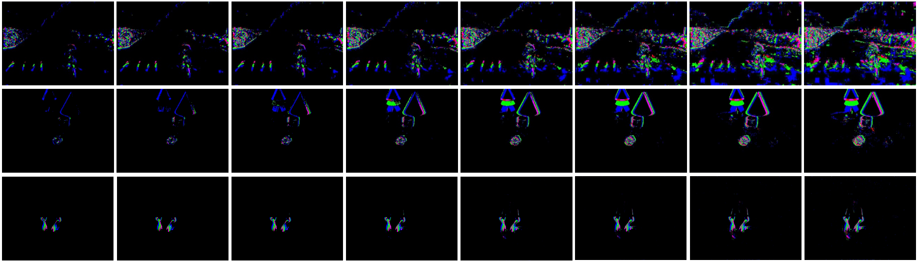
Taylor frames indicate motion strengths and directions.



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

Qualitative results (cont)

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

Quantitative results

	Model	Pretrain	Input	HMDB-51	CATER		MPII
					static	moving	
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 2: Evaluations on HMDB-51, CATER and MPII.

Quantitative results (cont.)

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

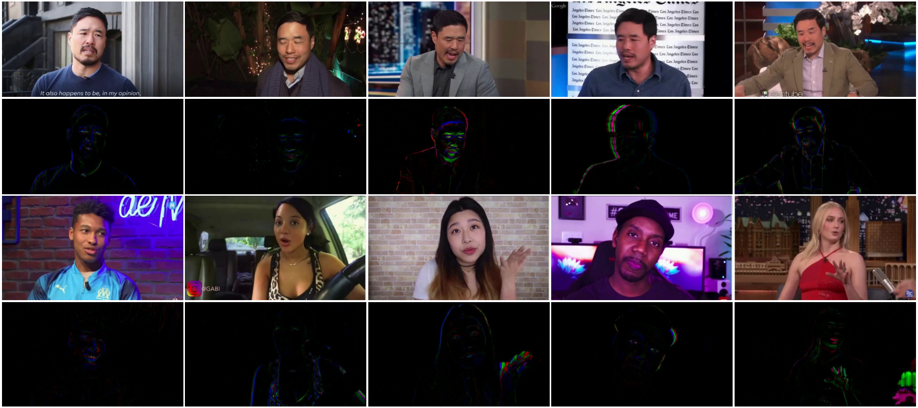
Table 3: Evaluations on Kinetics (K400 / K600) and Something-Something v2 (SSv2).

Model	Input	NTU-60		NTU-120		K-Skel
		X-Sub	X-View	X-Sub	X-Set	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 4: Comparing Taylor-transformed skeletons with original skeletons on NTU-60, NTU-120 and Kinetics-Skeleton (K-Skel).

Privacy-preserving

Taylor videos are able to remove distinct facial features of individuals compared to RGB videos. For more details, please refer to our paper[1].



[1] Wang, L., Xiu, Y., Gedeon, T., and Zheng, L.(2024). **Taylor Videos for Action Recognition**. In *ICML*.

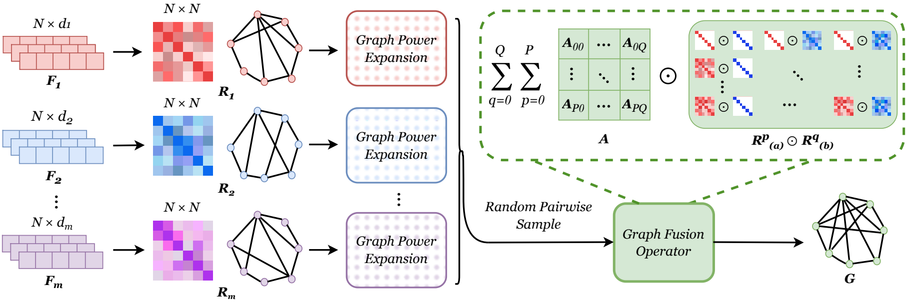
A Feature Fusion Framework: Learnable Expansion of Graph Operators

Motivation and key ideas

Whole fruits **Sliced fruit** **Fruit salad** **Mixed juice**

- Combining fruit slices resembles traditional early fusion methods, where features are concatenated but remain largely independent of each other
- We might cut the fruit into smaller pieces and mix them further, but the distinct flavors persist. This reflects late fusion methods, which combine outputs from separately trained models on different modalities
- While some integration occurs, the deeper interactions between the features are still missing, just as the flavors in the salad remain separate.
- Fruit mixer thoroughly blends the fruits, creating a smooth, unified mixture where each flavor enhances the whole. This blending captures the essence of feature fusion.

Motivation and key ideas (cont.)



- Text, images, and videos can be used to extract various *unit-level* features, ranging from word- and paragraph-level to patch-, clip-, frame-, cube-, or token-level, using pre-trained models.
- Relationship graph of unit-level features
- Heterogeneous features are transformed into a homogeneous graph space by modeling pairwise relationships among unit-level features, such as similarities, distances, or other relevant metrics.

Motivation and key ideas (cont.)

Consider two distinct relationship graphs $\boldsymbol{R}_{(a)}$ and $\boldsymbol{R}_{(b)}$. We construct a sequence of graph powers for each model or modality:

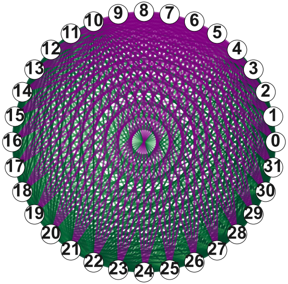
$$\begin{cases} \boldsymbol{\mathcal{G}}_{(a)} = \left[\boldsymbol{R}_{(a)}^0, \boldsymbol{R}_{(a)}^1, ..., \boldsymbol{R}_{(a)}^P\right] \in \mathbb{R}^{N\times N\times (P+1)} \\ \boldsymbol{\mathcal{G}}_{(b)} = \left[\boldsymbol{R}_{(b)}^0, \boldsymbol{R}_{(b)}^1, ..., \boldsymbol{R}_{(b)}^Q\right] \in \mathbb{R}^{N\times N\times (Q+1)} \end{cases}, \tag{2}$$

Mathematically, the graph fusion is expressed as follows:

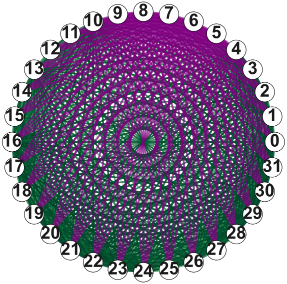
$$\begin{aligned} \boldsymbol{G} &= \boldsymbol{\mathcal{G}}_{(a)} \circledast \mathbf{A} \circledast \boldsymbol{\mathcal{G}}_{(b)}^\intercal \\ &= \sum_{q=0}^{Q}\sum_{p=0}^{P} \boldsymbol{R}_{(a)}^p a_p \odot \boldsymbol{R}_{(b)}^q b_q = \sum_{q=0}^{Q}\sum_{p=0}^{P} a_p b_q \left(\boldsymbol{R}_{(a)}^p \odot \boldsymbol{R}_{(b)}^q\right), \end{aligned} \tag{3}$$

where $\boldsymbol{a} = [a_p]_{p\in\mathcal{I}_{(P+1)}}$ and $\boldsymbol{b} = [b_q]_{q\in\mathcal{I}_{(Q+1)}}$ are the modality graph power selectors, and $\mathbf{A} = \boldsymbol{a}\otimes\boldsymbol{b} \in \mathbb{R}^{(P+1)\times(Q+1)}$, with $\otimes$ representing the outer product.

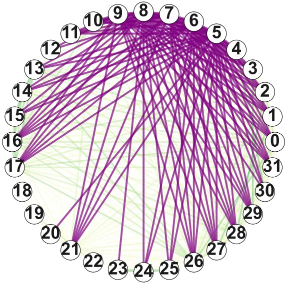
Qualitative results



(a) I3D visual features



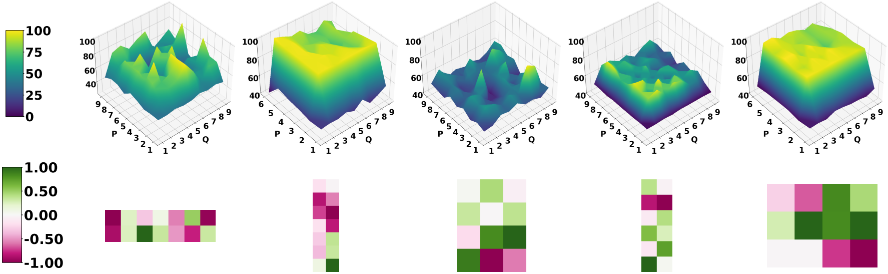
(b) SimCSE Text embeddings



(c) Fused relationship graph

- The graphs are constructed using cosine similarity to represent relationships among features.
- In each graph, nodes represent clip-level (or unit-level) features, with numbers indicating the sequence order of the video clips. Edges, shown in green, represent cosine similarity between features, with darker shades indicating stronger connections.
- Anomaly nodes and their connections are highlighted in purple (*e.g.*, the connection from node 4 to 10).

Qualitative results (cont.)



- (*Top row*): The effects of P (for visual feature) and Q (for text feature) in the learnable graph operator.
- (*Bottom row*): The learned optimal $\mathbf{A}$ for (*from left to right*) UCSD Ped2, ShanghaiTech, CUHK Avenue, Street Scene, and joint training on both UCSD Ped2 and ShanghaiTech.

Quantitative results

Table 5: Experimental results on feature-level and graph-level fusion across four video anomaly detection datasets, including single-modality comparisons. Graph-level single-modality and traditional methods use similarity graph representations for anomaly detection.

		UCSD Ped2	ShanghaiTech	CUHK Avenue	Street Scene
Feature-level	I3D visual	78.90	95.87	37.25	74.53
	Text only	80.02	83.39	65.19	69.34
	Concatenation	86.72	96.07	43.22	75.42
	Addition	86.20	95.77	57.44	75.05
	Product	62.72	94.15	32.04	75.59
	MTN fusion	92.80	96.37	62.06	71.50
Graph-level	I3D visual	68.89	69.88	58.72	49.12
	Text only	43.03	85.59	42.36	55.27
	Concatenation	63.45	88.68	50.09	48.97
	Addition	57.88	44.07	40.24	57.18
	Product	43.07	86.49	44.34	66.52
	EGO (ours)	**93.23**	**97.26**	**83.10**	**77.61**

Quantitative results (cont.)

Table 6: Comparison of Multi-scale Temporal Network (MTN) fusion (feature-level) and EGO fusion (graph-level). ShanghaiTech (ShT) is used for multi-representational and multi-modality fusion, while UCSD Ped2 (Ped2) and ShT are used for multi-domain fusion. Unlike MTN, which fuses two features at a time, EGO fusion enables simultaneous fusion of multiple features for greater flexibility. Training times for one epoch (in seconds) with a batch size of 32 on an Nvidia RTX 4070 GPU are also reported, with model sizes indicated in blue next to their respective models.

	Train	Test	MTN [29.0M]		**EGO** [0.091M]	
			AUC	Time	AUC	Time
Multi-represent.	I3D + C3D	I3D + C3D	**89.25**	13.6	87.17	**7.8**
	I3D + SwinT	I3D + SwinT	88.80	9.7	**89.85**	**4.9**
	C3D + SwinT	C3D + SwinT	84.45	12.0	**85.52**	**5.7**
	I3D + C3D + SwinT	I3D + C3D + SwinT	N/A	-	**95.38**	**9.0**
Multi-modality	Visual + Text	Visual + Text	96.37		**97.26**	
	Visual + Pose	Visual + Pose	95.48		**96.04**	
	Text + Pose	Text + Pose	94.49		**95.77**	
	Visual + Text + Pose	Visual + Text + Pose	N/A		**97.79**	
Multi-domain	Ped2 + ShT	Ped2 only	56.21		**58.30**	
		ShT only	**96.04**		95.10	
		Ped2 + ShT	**94.60**		92.11	

Robustness and Cross-Dataset Generalization

Table 7: Performance of EGO in visual and text fusion under **varying noise conditions** on text features using the ShanghaiTech dataset.

Condition	Original	10% Noise	30% Noise	50% Noise
Train on Noisy, Test on Clean	97.26	96.01	95.98	95.58
Train on Clean, Test on Noisy	97.26	95.96	95.86	95.62
Train on Noisy, Test on Noisy	97.26	95.92	95.76	94.86

Table 8: EGO performance on **different feature combinations**.

Feature Combination	EGO
I3D + SwinT	89.85
I3D + C3D	87.17
SwinT + C3D	85.52
I3D + SwinT + C3D	95.38

Table 9: Comparison of MTN fusion and EGO fusion performance in **cross-dataset evaluation**. Both models are trained on the ShanghaiTech dataset and evaluated on the UCSD Ped2, CUHK Avenue, Street Scene, XD-Violence, and UCF-Crime datasets.

Dataset	UCSD Ped2	CUHK Avenue	Street Scene	XD-Violence	UCF-Crime
MTN fusion	50.49	46.99	28.94	29.65	35.08
EGO (ours)	48.03	49.35	36.76	30.52	57.84

References and Further Reading

References and Further Reading

†: Corresponding author.

- **Lei Wang**†, Xiuyuan Yuan, Tom Gedeon, and Liang Zheng. “Taylor Videos for Action Recognition.” *ICML*. 2024. A*
- Liyun Zhu, **Lei Wang**†, Arjun Raj, Tom Gedeon, and Chen Chen. “Advancing Video Anomaly Detection: A Concise Review and a New Dataset.” *NeurIPS D&B Track*. 2024. A*
- Dexuan Ding, **Lei Wang**†, Liyun Zhu, Tom Gedeon, Piotr Koniusz. “Learnable Expansion of Graph Operators for Multi-Modal Feature Fusion.” *ICLR*, 2025. A*
- Qixiang Chen, **Lei Wang**†, Piotr Koniusz and Tom Gedeon. “Motion meets Attention: Video Motion Prompts.” *ACML*. 2024. Oral [5.67% acceptance rate (20.11% overall acceptance rate)]
- Xi Ding, **Lei Wang**†. “The Journey of Action Recognition.” *Companion Proceedings of the ACM Web Conference (WWW Companion)*, 2025. Oral
- Xi Ding, **Lei Wang**†. “Do Language Models Understand Time?” *Companion Proceedings of the ACM Web Conference (WWW Companion)*, 2025. Oral

Q&A — Thank you!