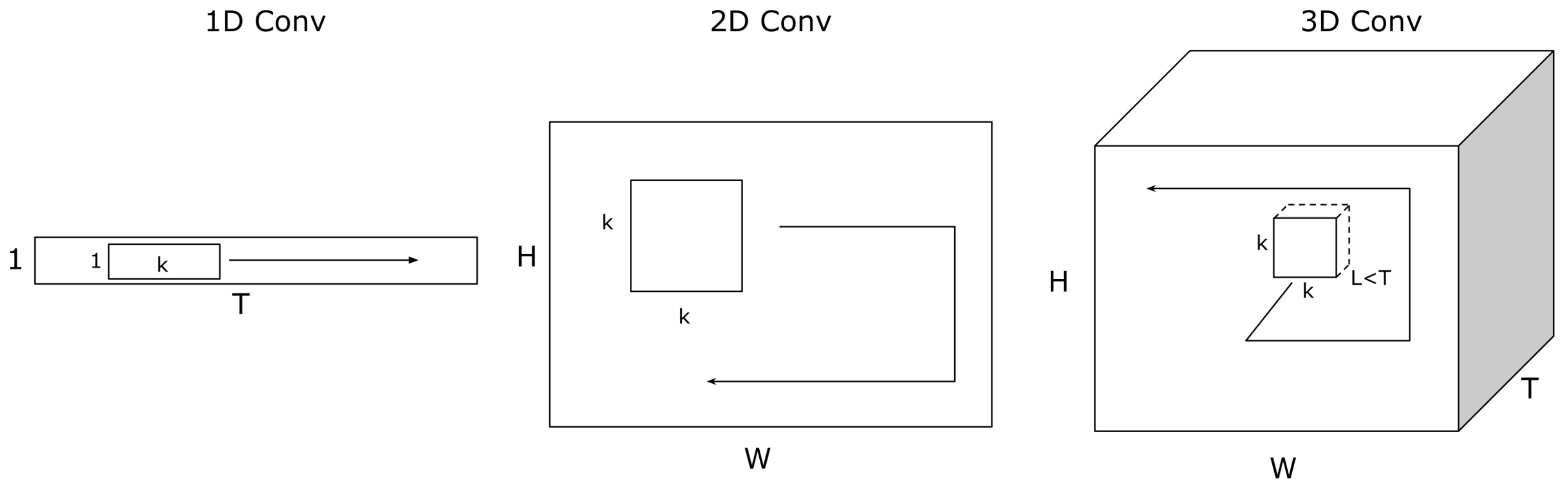
Temporal Gaussian Mixture Layer for Videos

AJ Piergiovanni and Michel S. Ryoo

Indiana University

Motivation – Video Representation Learning

- Learning good video representations has many applications
  - Robot perception, activity recognition, smart cities, sports analysis
- Videos are high-dimensional spatio-temporal data, abstracting representations is critical for many tasks
- Standard methods use CNNs with temporal convolution (e.g., 1D or 3D convolution)



Temporal Information is Needed

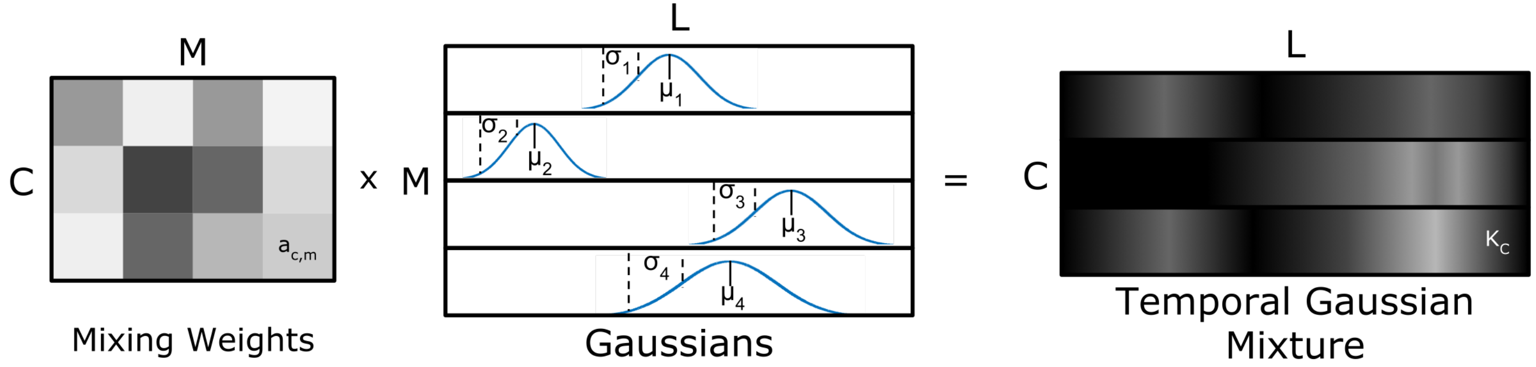
- Standard CNNs only capture short-term information
  - 2D CNNs use a single frame
  - 3D CNNs capture only 2-3 seconds
- Short clips can be ambiguous

Temporal Information is Needed

- Standard CNNs only capture short-term information
- Short clips can be ambiguous
- Extending 3D/1D conv to longer durations leads to many parameters and poor performance

Temporal Gaussian Mixture Layer

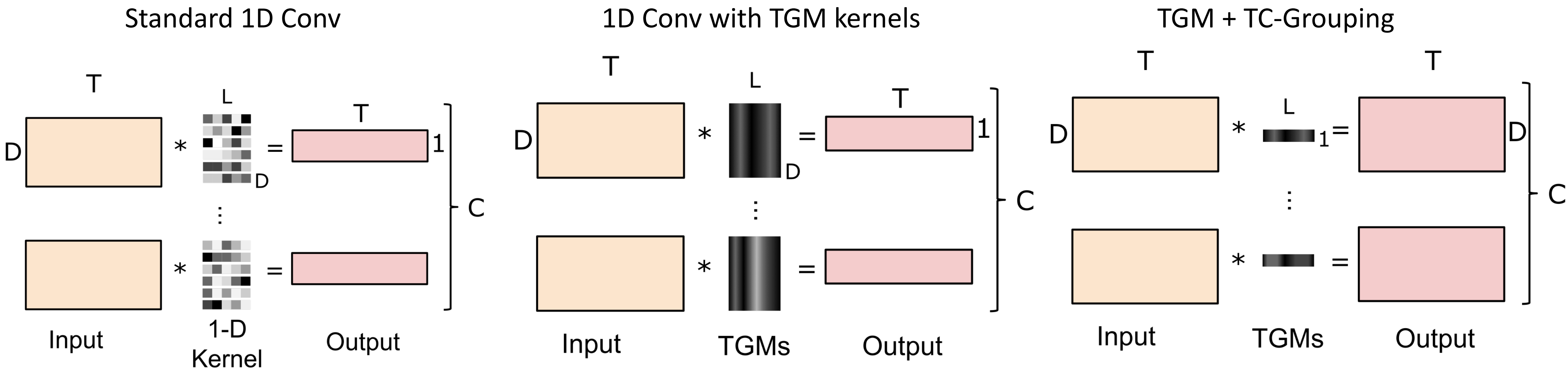
- Can learn longer-term temporal structures without increasing parameters
- Learns a set of Gaussians and mixing weights which generates the temporal convolutional kernel



$$a_{i,m} = \frac{\exp \omega_{i,m}}{\sum_j \exp \omega_{i,j}} \qquad \hat{K}_{m,l} = \frac{1}{Z} \exp - \frac{(l - \mu_m)^2}{2\sigma_m^2} \qquad K_i = \sum_m a_{i,m} \hat{K}_i$$

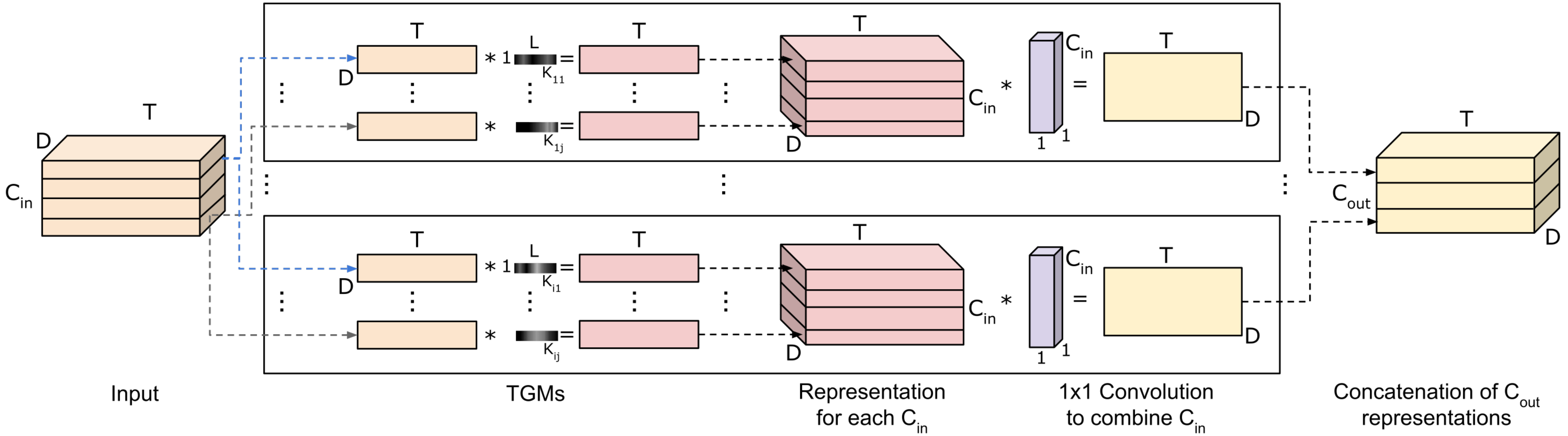
Using TGMs

- Can apply TGM as standard 1D convolution or as grouped 2D convolution
    - Loses some information when combining the base CNN channels



Temporal Channel Grouped Convolution

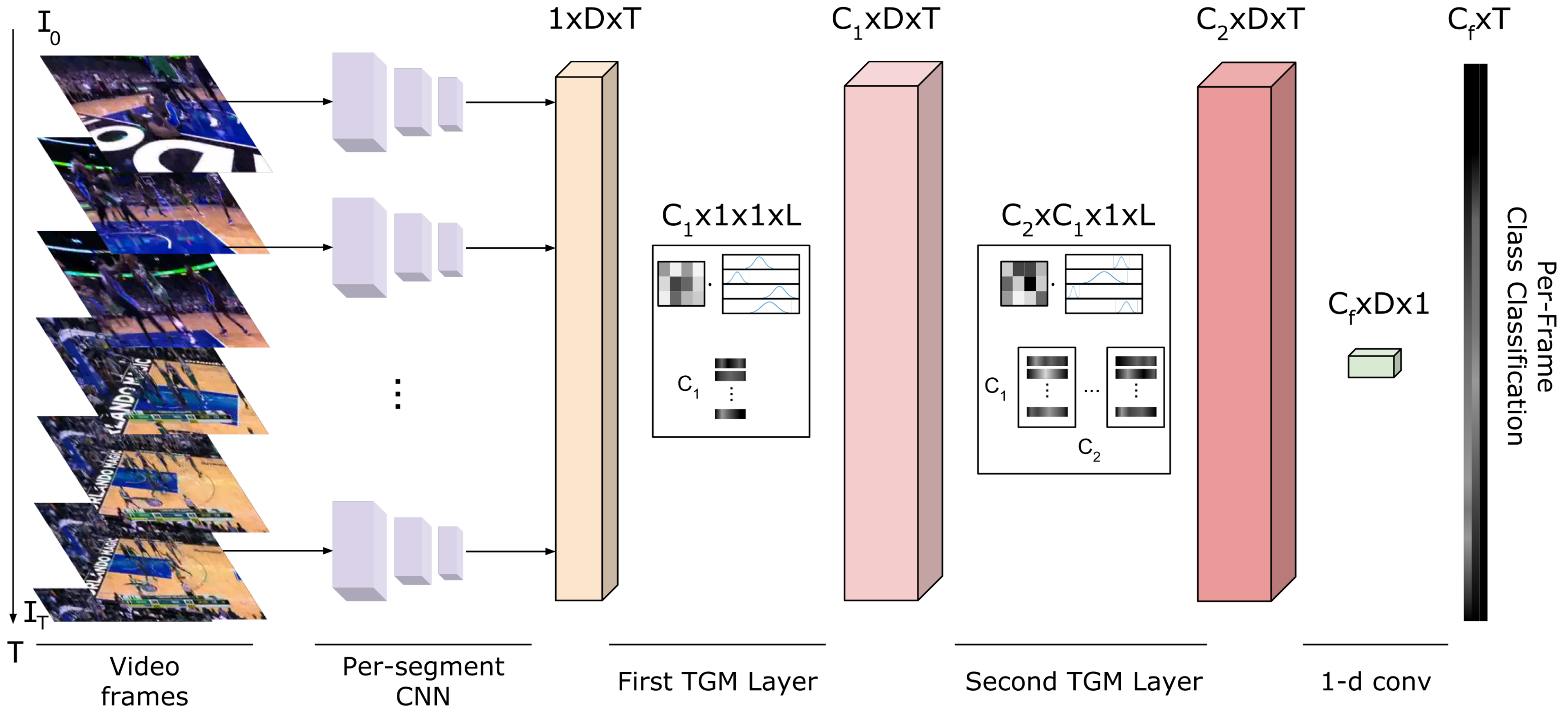
- TC-Grouping **adds** a new temporal channel axis
  - Allows for learning of different temporal structures with base CNN feature channels



$$s_i = G_i * w_i = (f_j * K_{i,j}) * w_i, \;\; S = [s_1, s_2 \ldots, s_{C_{out}}]$$

Activity Detection with TGMs

- Applies base CNN, followed by TGMs to learn longer-term temporal structure, followed by a classification layer.



Fewer Parameters

Model	# of parameters
LSTM	10.5M
1 Temporal Conv	10.5M
3 Temporal Conv	31.5M
1 TGM Layer	10K
3 TGM Layers	100K
5 TGM Layers	200K

LSTMs and 1D Conv with fewer parameters leads to nearly random performance.

Model	mAP
LSTM with 100k parameters	6.5
Temporal Conv. with 100k parameters	7.3
TGM with random temporal filters	34.5
TGM with fixed Gaussians	38.5
Full TGM	**44.3**

Fewer Parameters

Model	# of parameters
LSTM	10.5M
1 Temporal Conv	10.5M
3 Temporal Conv	31.5M
1 TGM Layer	10K
3 TGM Layers	100K
5 TGM Layers	200K

LSTMs and 1D Conv with fewer parameters leads to nearly random performance.

Model	mAP
LSTM with 100k parameters	6.5
Temporal Conv. with 100k parameters	7.3
TGM with random temporal filters	34.5
TGM with fixed Gaussians	38.5
Full TGM	**44.3**

Stacking 1D conv reduces performance, but stacking TGMs is beneficial

Model	Spatial	Temporal	Two-stream
Random	13.4	13.4	13.4
I3D	33.8	35.1	34.2
I3D + LSTM	36.2	37.3	39.4
I3D + temporal conv	37.3	38.6	39.9
I3D + 3 temporal conv	32.4	34.6	35.6
I3D + 1 TGM	35.5	37.5	38.5
I3D + 3 TGM	36.5	38.4	40.1

Results on MultiTHUMOS

	mAP
Two-stream (Yeung et al., 2015)	27.6
Two-stream + LSTM (Yeung et al., 2015)	28.1
Multi-LSTM (Yeung et al., 2015)	29.6
Predictive-corrective (Dave et al., 2017)	29.7
SSN (Zhao et al., 2017)	30.3
I3D baseline	29.7
I3D + LSTM	29.9
I3D + temporal pyramid	31.2
I3D + super-events (Piergiovanni & Ryoo, 2018b)	36.4
I3D + our TGMs	44.3
I3D + super-events + our TGMs	**46.4**

Ground Truth Baseline Super-Events TGM Full

Results on Charades

	mAP
Predictive-corrective (Dave et al., 2017)	8.9
Two-stream (Sigurdsson et al., 2016a)	8.94
Two-stream+LSTM (Sigurdsson et al., 2016a)	9.6
R-C3D (Xu et al., 2017)	12.7
Sigurdsson et al. (Sigurdsson et al., 2016a)	12.8
SSN (Zhao et al., 2017)	16.4
I3D baseline	17.2
I3D + 3 temporal conv. layers ($L = 5$)	17.5
I3D + 3 temporal conv. layers ($L = 30$)	12.5
I3D + LSTM	18.1
I3D + fixed temporal pyramid	18.2
I3D + super-events (Piergiovanni & Ryoo, 2018b)	19.4
I3D + 3 TGMs ($L = 5$)	20.6
I3D + 3 TGMs ($L = 30$)	21.5
I3D + 3 TGMs ($L = 5$) + super-events	21.8
I3D + 3 TGMs ($L = 30$) + super-events	**22.3**

Ground Truth Baseline Super-Events TGM Full

Increasing temporal resolution

- Increasing 1-D conv size reduces performance
- Increasing TGMs adds no parameters, improves performance and focuses on important intervals

	MultiTHUMOS			Charades		
	1 Layer	3 Layers	1-D Conv	1 Layer	3 Layers	1-D Conv
I3D Baseline	22.3	-	-	15.3	-	-
$L = 3$	30.2	31.7	26.6	15.5	16.1	15.5
$L = 5$	32.5	**37.2**	28.3	15.7	17.8	16.3
$L = 10$	34.5	35.4	31.7	16.1	18.2	16.6
$L = 15$	**36.1**	34.1	32.5	17.5	18.6	16.8
$L = 30$	32.5	33.9	26.5	18.1	**18.9**	12.1
$L = 50$	32.1	33.7	15.4	**18.3**	18.8	6.7

Thank you

Please visit our poster #149 for more details

Code and models:

https://github.com/piergiaj/tgm-icml19