Temporal Transformer Networks: Joint Learning of Invariant and Discriminative Time Warping

Suhas Lohit, Qiao Wang, Pavan Turaga

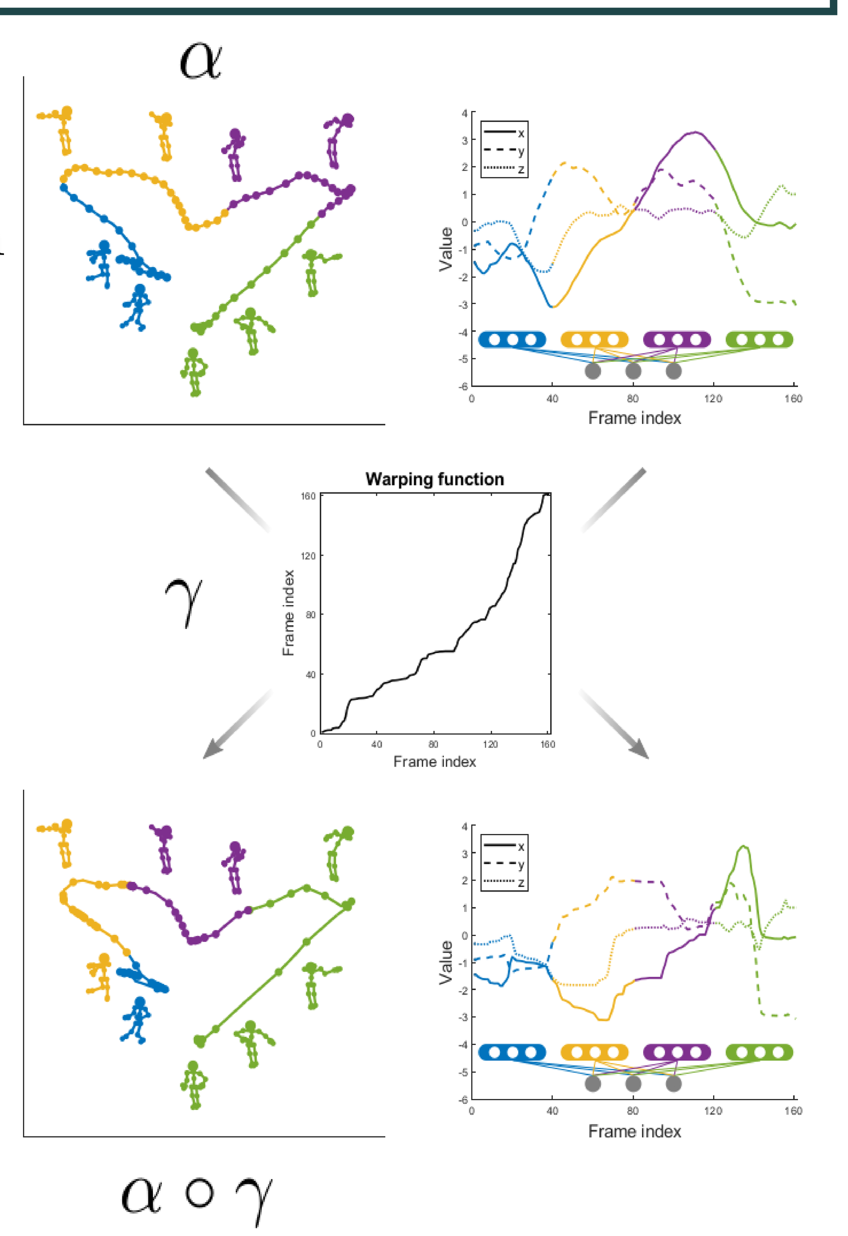
Geometric Media Lab, Arizona State University, Tempe, AZ

`{slohit, qiao.wang, pturaga}@asu.edu`

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Rate-invariant action recognition

- Invariance to execution rate is important for time-series classification such as human action recognition
- Conventional neural networks are not designed to guarantee rate-invariance
- We design a specialized module – the temporal transformer – which provides improved discrimination and invariance for time-series classification



Order-preserving diffeomorphisms

- Rate-modifying transforms are easily modeled using order-preserving diffeomorphisms, γ[1] :

$$\gamma(0) = 0, \gamma(1) = 1$$

$$\gamma(t_1) < \gamma(t_2), \text{ if } t_1 < t_2$$

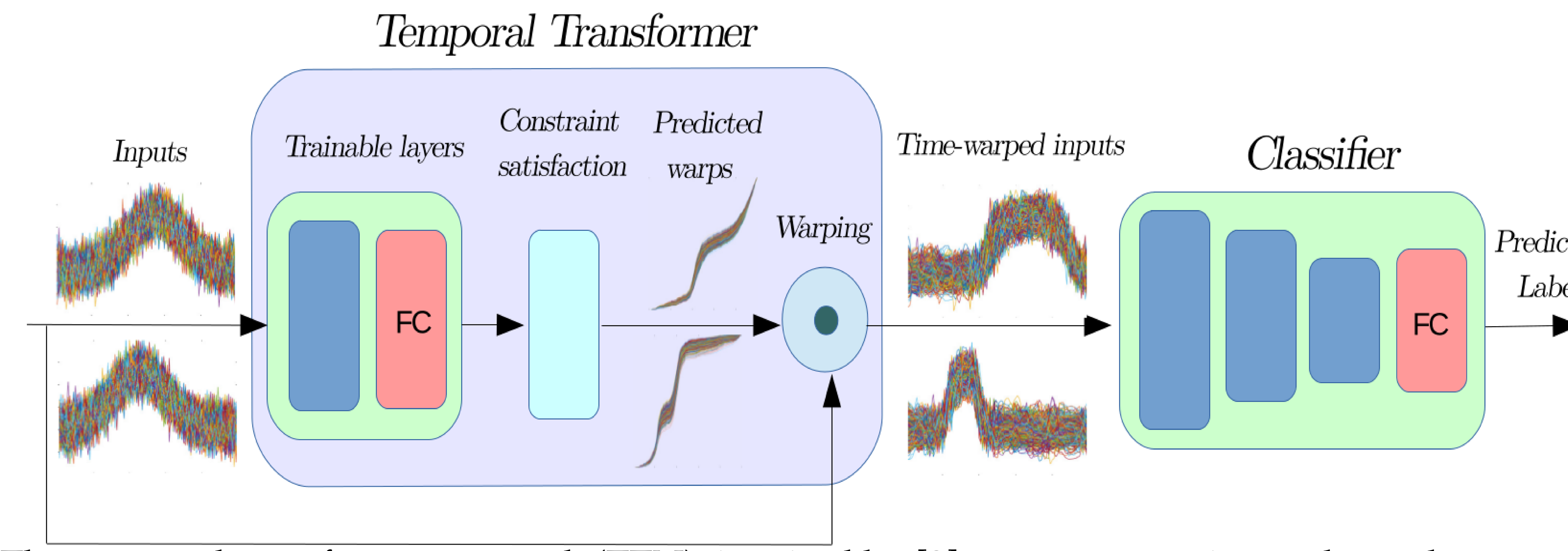
- γ has the properties of a cumulative distribution function

$$\gamma(t) = \int_0^t \dot{\gamma}(t)dt$$

$$\int_0^1 \dot{\gamma}(t)dt = \gamma(1) - \gamma(0) = 1$$

- This is a non-parametric set of transforms with order-preserving and end-point constraints, different from what is studied in literature [2]

Differentiable module for warping time



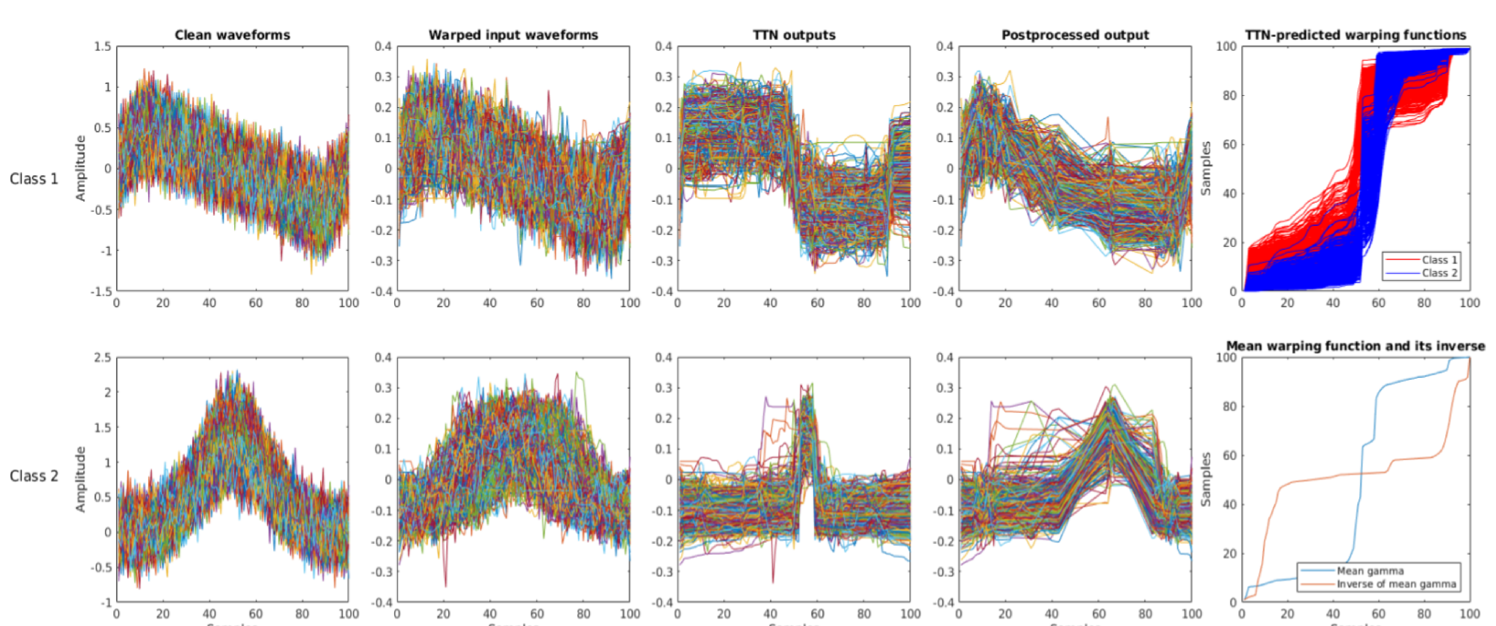
- The temporal transformer network (TTN), inspired by [2], generates an input-dependent γ, which is used to warp the input time series before classification so as to maximize recognition accuracy
- Constraint satisfaction ensures that the output of TTN is an order-preserving diffeomorphism:

$$\dot{\gamma} = \frac{\mathbf{v}}{\|\mathbf{v}\|} \odot \frac{\mathbf{v}}{\|\mathbf{v}\|}, \quad \text{and} \quad \gamma(t) = T \cdot \sum_{i=1}^{t} \dot{\gamma}(i)$$

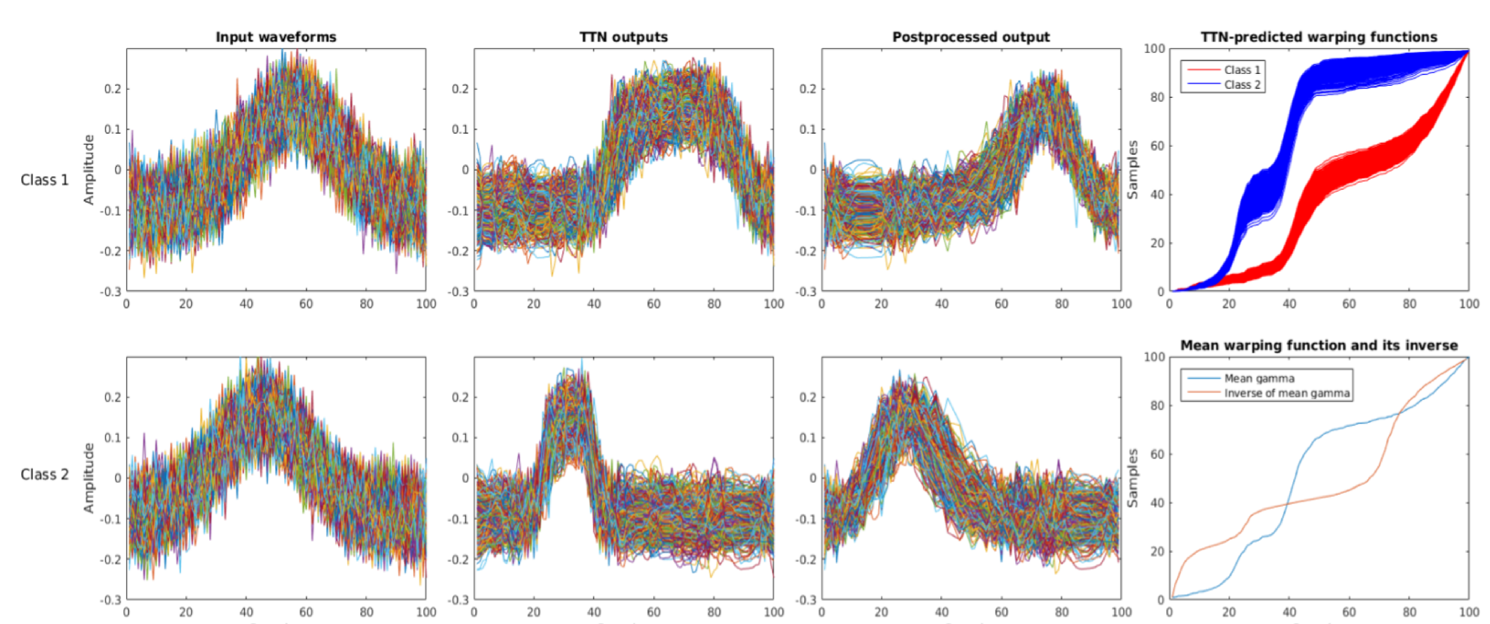
- Warping is performed using linear interpolation which is differentiable

Experiments on synthetic data

Demonstrating rate-invariance properties of TTN



Demonstrating class-discriminative properties of TTN

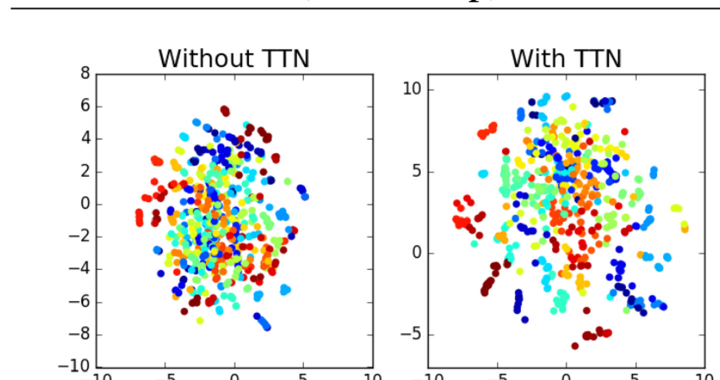


Experiments in skeletal action recognition

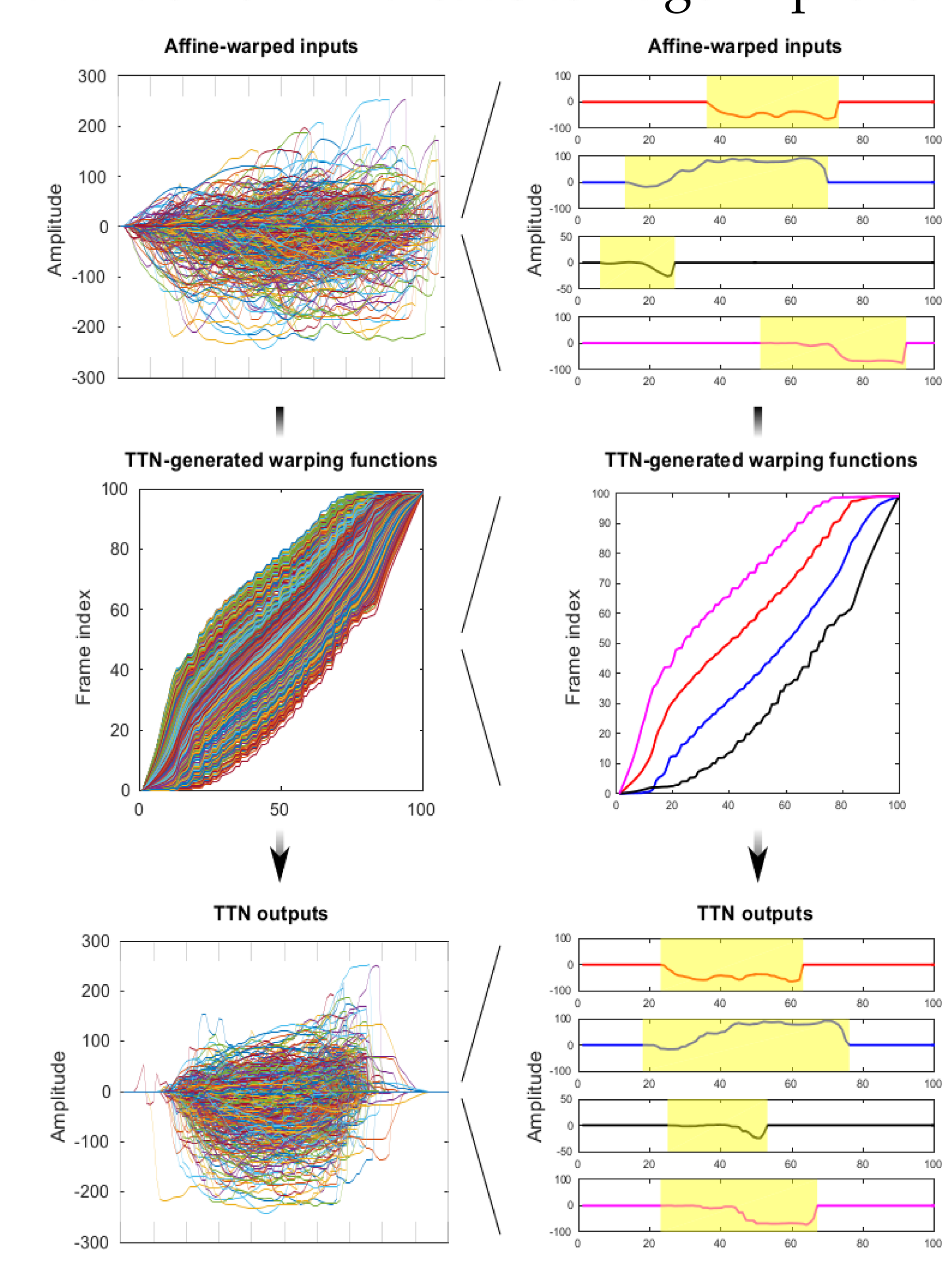
ICL First-Person Hand Action dataset [3]:

- Mocap dataset with 600 training and 575 testing 3D pose seqeunces of 26 actions.
- We experiment with both 1-layer TCN and 2-layer LSTM. In both cases, adding the TTN (3 FC layers) improves performance significantly

Method	Accuracy (%)
2-layer LSTM	76.17
2-layer LSTM + TTN	**78.43**
TCN-16	76.28 $\pm$ 0.29
TCN-16 + TTN	**80.14 $\pm$ 0.33**
TCN-64	79.10 $\pm$ 0.76
TCN-64 + TTN	**81.32 $\pm$ 0.36**
TCN-32	81.74 $\pm$ 0.27
TCN-32 + TTN	**82.75 $\pm$ 0.31**
TCN-32 (affine warp)	70.43
TCN-32 + TTN (affine warp)	**78.26**



- In the presence of affine warp distortion, addition of TTN leads to huge improvements



NTU RGB-D dataset [4]:

- A large Kinect dataset with 56000 human action sequences, 60 actions by 45 subjects
- We use TCN with 10 conv layers as the base classifier
- Adding the TTN (2 conv + 3 FC layers) module improves recognition performance

Method	CS (%)	CV (%)
Lie Groups	50.08	52.76
FTP Dynamic Skeletons	60.23	65.22
HBRNN	59.07	63.97
2-layer part-LSTM	62.93	70.27
STA-LSTM	73.40	81.20
VA-LSTM	79.40	87.60
STA-GCN	***81.50***	***88.30***
TCN	76.54	83.98
TCN + TTN	**77.55**	**84.25**

CS : Cross Subject

CV : Cross View

[1] Srivastava, Anuj, and Eric P. Klassen. Functional and shape data analysis. New York: Springer, 2016.
[2] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. "Spatial transformer networks." Advances in neural information processing systems. 2015.
[3] Shahroudy, Amir, et al. "NTU RGB+ D: A large scale dataset for 3D human activity analysis." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016
[4] Garcia-Hernando, Guillermo, et al. "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

https://github.com/suhaslohit/TTN