Diffeomorphic Temporal Alignment Nets

Ron Shapira Weber[1], Matan Eyal[1], Nicki Skafte Detlefsen[2], Oren Shriki[1] and Oren Freifeld[1]

1. Ben-Gurion University
2. Technical University of Denmark

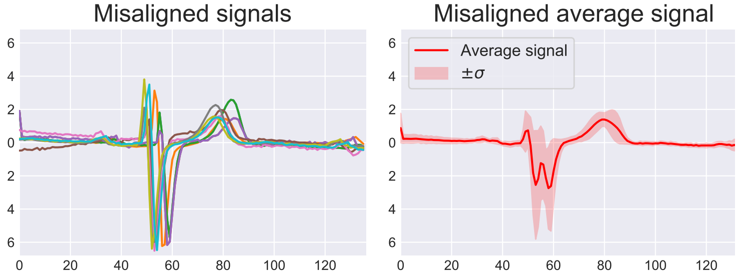
October 2019

Author's summary (NeurIPS 2019)

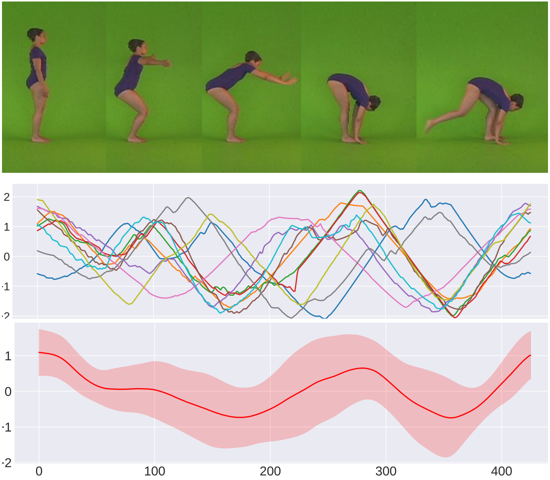
Outline

Problem Formulation - Time-Series Joint Alignment

- Time-series data often presents a significant amount of nonlinear misalignment.



Example - Human Activity Recognition via Repeated Trials



Yoga dataset of the UCR archive (Chen., 2015)

Preliminaries - Spatial Transformer Nets

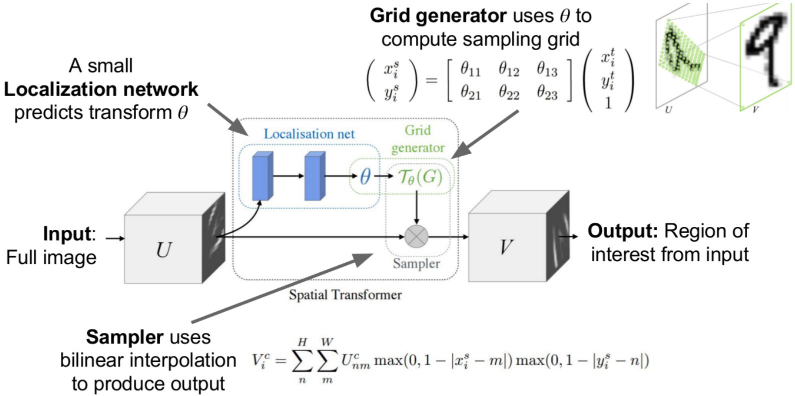


Figure taken from: Jaderberg et al. NIPS (2015) and Manjunath Bhat Medium post.

Preliminaries - Spatial/Temporal Transformer Nets

- Temporal Transformer Nets (TTN), the time-series analog of STN.
- In more detail, let $\boldsymbol{U}$ denote the input of the TT layer. Its output consists of $\boldsymbol{\theta} = f_{\mathrm{loc}}(\boldsymbol{U})$ and $\boldsymbol{V} = \boldsymbol{U} \circ T^{\boldsymbol{\theta}}$ (the latter, *i.e.*, the warped signal, is what is being passed downstream the TTN), where $T^{\boldsymbol{\theta}} \in \mathcal{T}$ is a 1D warp parameterized by $\boldsymbol{\theta}$. The function $f_{\mathrm{loc}} : \boldsymbol{U} \mapsto \boldsymbol{\theta}$ is itself a neural net called the localization net.

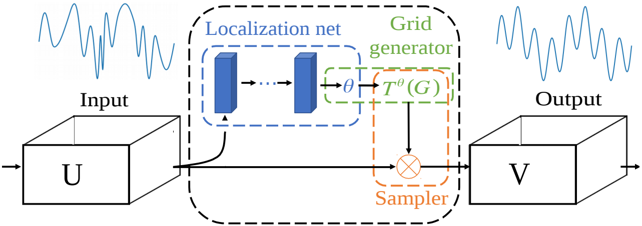
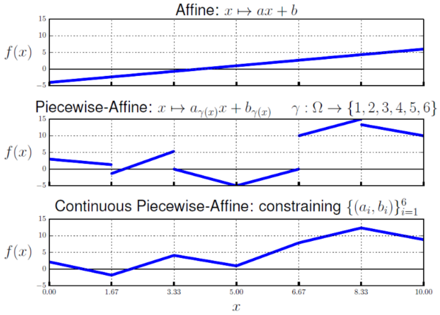


Figure adapted from: Skafte et al. CVPR (2018)

Preliminaries - CPAB

- The diffeomorphism family we use in this paper is CPAB (Freifeld et al. (ICCV, 2015; PAMI, 2017)).
- The name CPAB, short for CPA-Based, is due to the fact that these warps are based on Continuous Piecewise-Affine (CPA) velocity fields. The term "piecewise" is w.r.t. a partition, denoted by Ω, of the signal's domain into subintervals.



Preliminaries - CPAB

- Let $\mathcal{V}$ denote the linear space of CPA velocity fields w.r.t. such a fixed Ω, let $d = \dim(\mathcal{V})$, and let $v^{\boldsymbol{\theta}} : \Omega \to \mathbb{R}$, a velocity field parametrized by $\boldsymbol{\theta} \in \mathbb{R}^d$, denote the generic element of $\mathcal{V}$, where $\boldsymbol{\theta}$ stands for the coefficient w.r.t. some basis of $\mathcal{V}$ is

$$\mathcal{T} \triangleq \{T^{\boldsymbol{\theta}} : x \mapsto \phi^{\boldsymbol{\theta}}(x;1) \text{ s.t. } \phi^{\boldsymbol{\theta}}(x;t) = x + \int_0^t v^{\boldsymbol{\theta}}(\phi^{\boldsymbol{\theta}}(x;\tau))\,\mathrm{d}\tau\}; \quad (1)$$

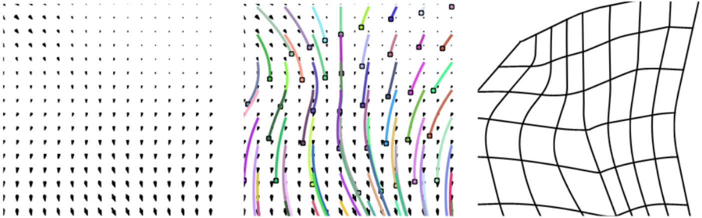
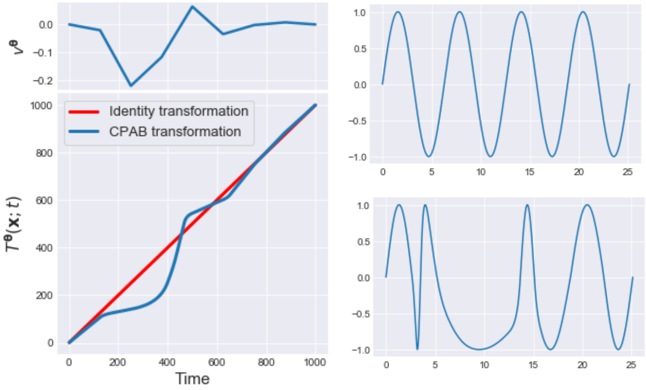


Figure taken from: Freifeld et al. PAMI (2017)

CPAB - 1D Example

- Left: An illustration of a CPAB warp (relative to the identity transformation) with its corresponding CPA velocity field (above). Right: a sine wave before (top) and after (bottom) being warped by the presented CPAB transformation.



Loss Function

- As the variance of the observed $(\boldsymbol{U}_i)_{i=1}^N$ is (at least partially) explained by the latent warps, $(T^{\boldsymbol{\theta}_i})_{i=1}^N$, we seek to minimize the empirical variance of the warped signals, $(\boldsymbol{V}_i)_{i=1}^N$. In other words, our data term in this setting is

$$F_{\text{data}}\left(\boldsymbol{w}, (\boldsymbol{U}_i)_{i=1}^N\right) \triangleq \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{V}_i(\boldsymbol{U}_i;\boldsymbol{w}) - \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{V}_j(\boldsymbol{U}_j;\boldsymbol{w})\right\|_{\ell_2}^2 \tag{2}$$

- For multi-class problems, our data term is the sum of the within-class variances:

$$F_{\text{data}}\left(\boldsymbol{w}, (\boldsymbol{U}_i)_{i=1}^N\right) \triangleq$$
$$\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i:z_i=k}\left\|\boldsymbol{V}_i\left(\boldsymbol{U}_i;\boldsymbol{w}\right) - \frac{1}{N_k}\sum_{j:z_j=k}\boldsymbol{V}_j(\boldsymbol{U}_j;\boldsymbol{w})\right\|_{\ell_2}^2$$

- where K is the number of classes, z_i takes values in $\{1,\ldots,K\}$.

Loss Function - Regularization

- In both the single- and multi-class cases, we also use a regularization term on the warp

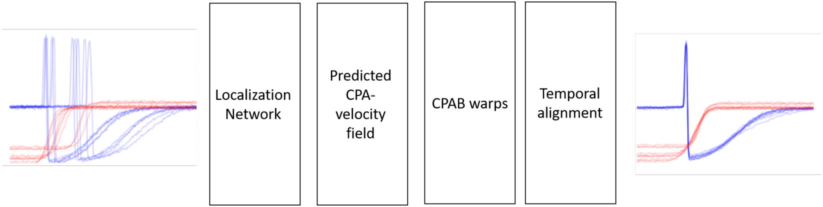
$$F_{\mathrm{reg}}(\boldsymbol{w},(\boldsymbol{U}_i)_{i=1}^N) = \sum\nolimits_{i=1}^{N}(\boldsymbol{\theta}_i(\boldsymbol{w},\boldsymbol{U}_i))^T\boldsymbol{\Sigma}_{\mathrm{CPA}}^{-1}\boldsymbol{\theta}_i(\boldsymbol{w},\boldsymbol{U}_i) \tag{3}$$

- Where $\boldsymbol{\Sigma}_{\mathrm{CPA}}$ is a CPA covariance matrix (proposed by Freifeld et al. (ICCV, 2015; PAMI, 2017) associated with a zero-mean Gaussian smoothness prior over CPA fields.
- $\boldsymbol{\Sigma}_{\mathrm{CPA}}$ has two parameters: λ_{var}, which controls the overall variance, and $\lambda_{\mathrm{smooth}}$, which controls the smoothness of the field.
- A small λ_{var} favors small warps (*i.e.*, close to the identity); similarly, the larger $\lambda_{\mathrm{smooth}}$ is, the more it favors CPA velocity fields that are almost purely affine.
- Our loss function, to be minimized over $\boldsymbol{w}$, is

$$F(\boldsymbol{w},(\boldsymbol{U}_i)_{i=1}^N) = F_{\mathrm{data}}(\boldsymbol{w},(\boldsymbol{U}_i)_{i=1}^N) + F_{\mathrm{reg}}(\boldsymbol{w},(\boldsymbol{U}_i)_{i=1}^N)\,. \tag{4}$$

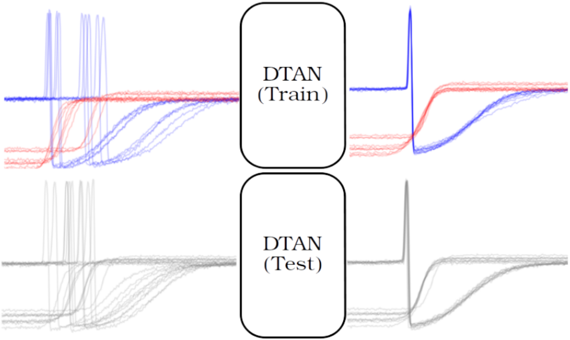
Diffeomorphic Temporal Alignment Nets

- During train, DTAN is set to minimize the joint alignment loss via temporal warping of its input-signals.
- For each batch, the alignment is within-class.



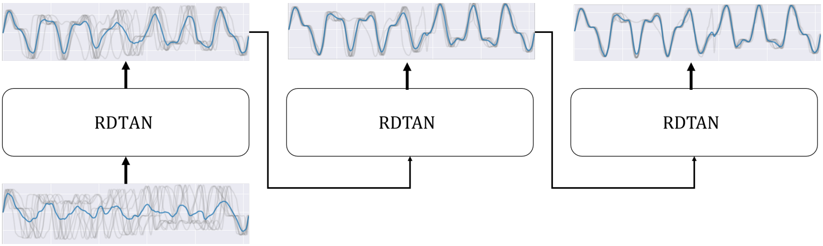
Diffeomorphic Temporal Alignment Nets

- At test time, the class labels are unknown.



Recurrents DTANs

- RDTAN - recurrent warps of the input signals.
- Increasing the number of recurrent warps does not increase the number of trainable parameters as the same localization network is used.



Experiments and Results

- We first explore DTAN time-series joint alignment of synthetic data. Unlike real-world data, here the latent warp signal is available for us.

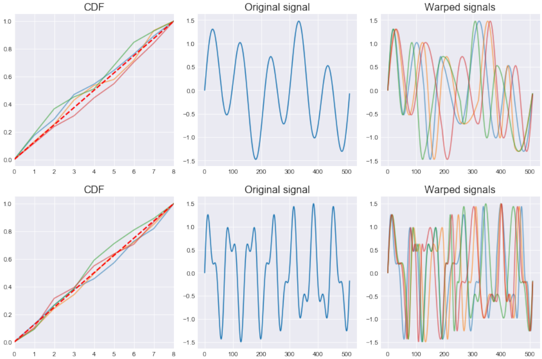
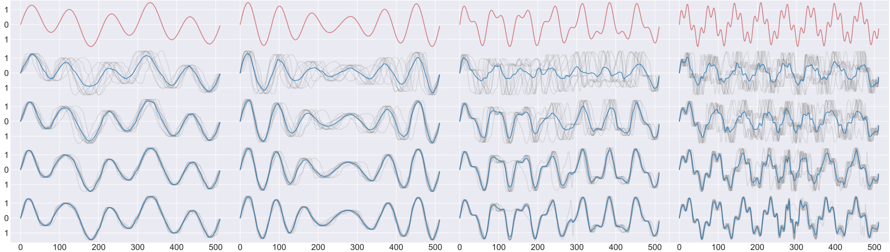


Figure: Synthetic data generation process. We create 4 latent signal (2 are presented here). We then sample CDFs from a Dirichlet distribution and use them as random warp.

R-DTAN Joint Alignment of Synthetic Data

- Here we see the output of R-DTAN at each recurrence of the synthetic **test data**.
- The original signals are in red. Each column represents a different class. Note that the same model was trained on all classes.



Synthetic Data - Variance Reduction

- DTAN exhibits comparable results in terms of variance reduction between the train and test sets.
- Increasing the number of applied warps via an R-DTAN (without increasing the number of learned parameters) further decreases the variance.

	Train set variance				Test set variance			
Dataset	Baseline	DTAN	R-DTAN2	R-DTAN4	Baseline	DTAN	R-DTAN2	R-DTAN4
Dir(32)	0.483	0.136	0.106	**0.088**	0.466	0.234	0.167	**0.130**
Dir(16)	0.522	0.240	0.162	**0.098**	0.514	0.332	0.24	**0.154**
Dir(8)	0.536	0.254	0.181	**0.122**	0.532	0.362	0.248	**0.183**

Synthetic Data - Timing

- DTAN joint alignment timing w.r.t. signal's length and size of the test-set (16 sets in total).

Alignment timing per test set (in [sec])				
length \ # of signals	10	10^2	10^3	10^4
64	0.003	0.003	0.007	0.109
128	0.003	0.004	0.012	0.211
256	0.014	0.038	0.042	0.455
512	0.003	0.007	0.084	0.660

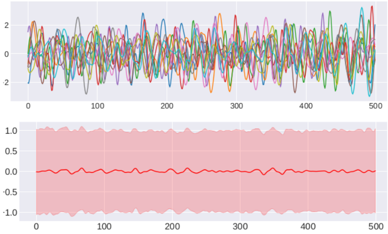
Joint Alignment of Real-World data

- The UCR time-series classification archive (Chen et al, 2015) contains 85 real-world datasets (we used 84).
- f_{loc} - standard 1D-CNN consisting of 3 conv-layers (1286464 filters per layer, respectively) each followed by a ReLU nonlinear activation function. $d = \dim(\boldsymbol{\theta}) = 32$.
- Below: The ECGFiveDays dataset. Each panel depicts 10 random sample from the train/test sets and their respective mean (shaded areas correspond to standard deviations).

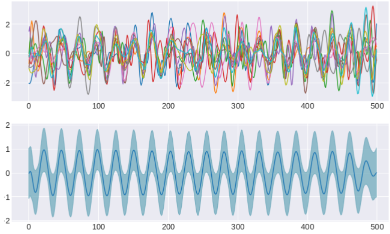
(a) Train (b) Test

Joint Alignment: More Results

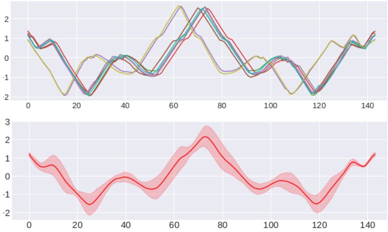
(a) FordA – misaligned



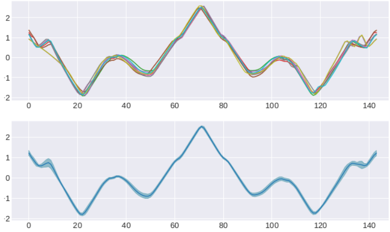
(b) FordA – aligned



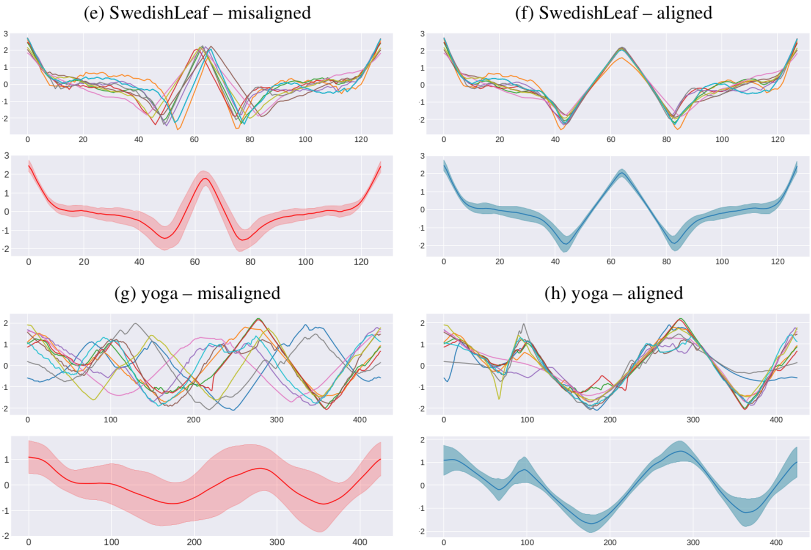
(c) Plane – misaligned



(d) Plane – aligned

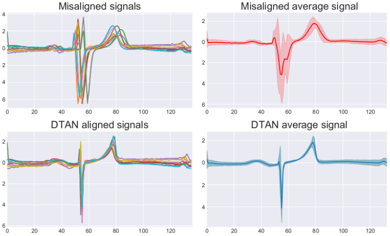


Joint Alignment: More Results

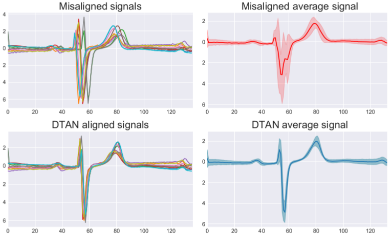


Regularization effect - W/O

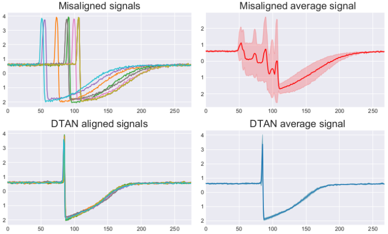
(c) ECGFiveDays - Without regularization



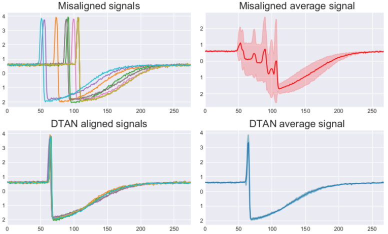
(d) With regularization



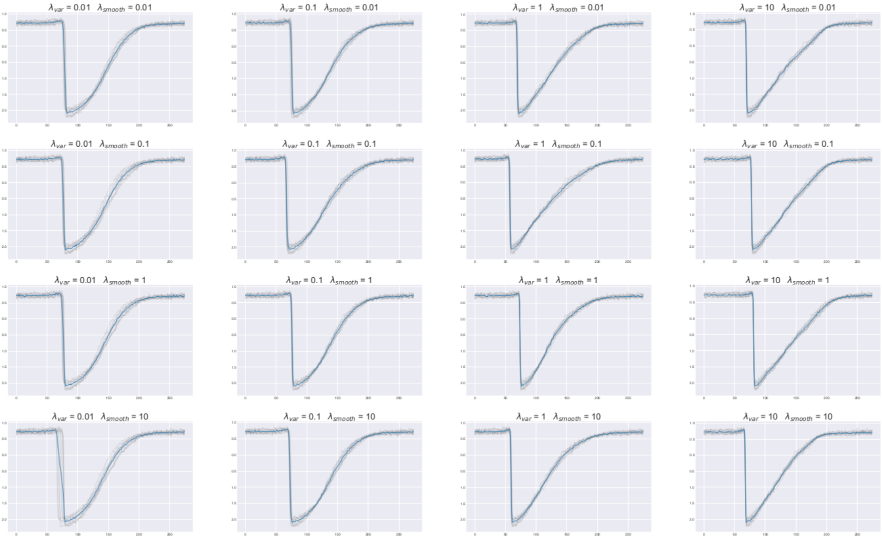
(e) Trace - Without regularization



(f) With regularization



Regularization effect - Different values



Nearest Centroid Classification (NCC) experiment

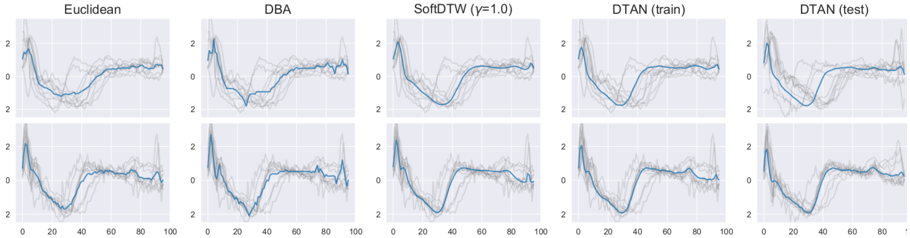
- Similar to 1-NN classifer - using each class average (centroid) instead of the entire train set.
- When using DTW distance, on-par with some of the state-of-the-art algorithms for time-series classification.
- Used here to evaluate the quality of DTAN's average signal.
- We compared DTAN NCC while using euclidean distance to DBA and SoftDTW (while measuring DTW).

NCC Results

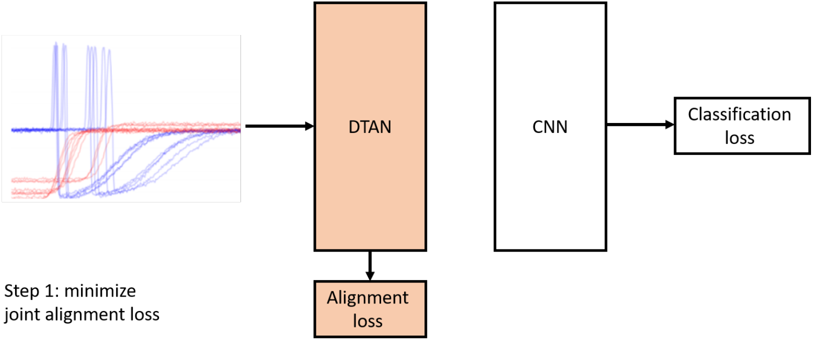
- Correct classification rates using NCC. Each point above the diagonal indicates an entire UCR archive dataset where DTAN achieved better (or no-worse) results than the competing method.

- DTANs test accuracy compared with: Euclidean (DTAN was better or no worse in 93% of the datasets), DBA (77%) and SoftDTW (62%).

Time-Series Averaging Comparison

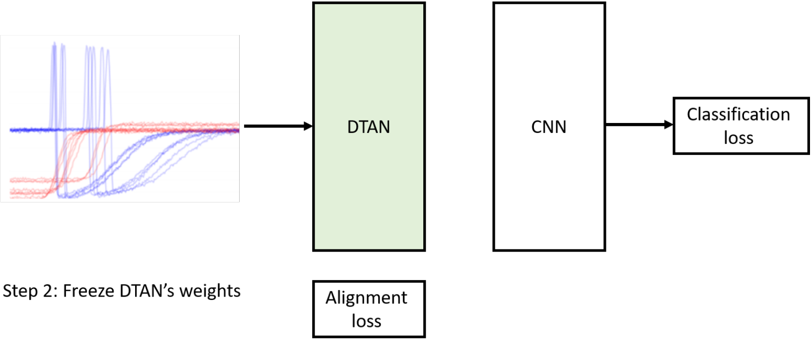


CNN Classification Experiment



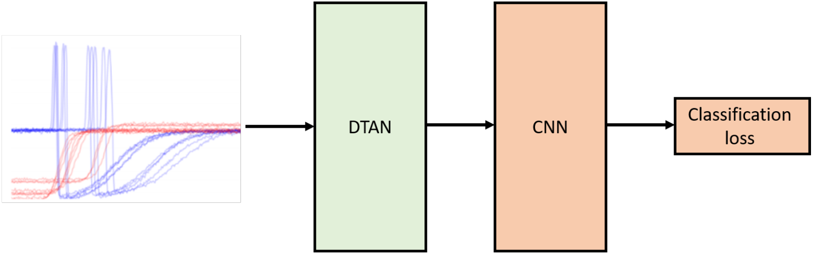
Step 1: minimize
joint alignment loss

CNN Classification Experiment



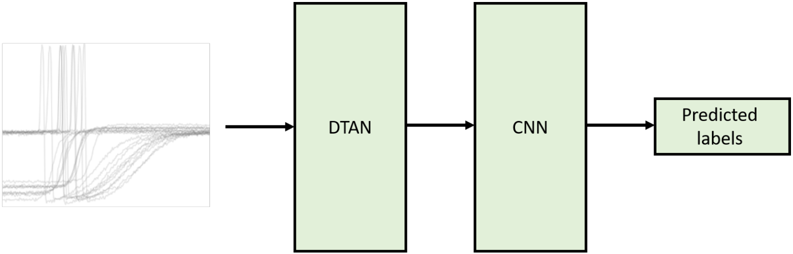
Step 2: Freeze DTAN's weights

CNN Classification Experiment

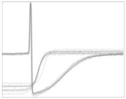


Step 3: Connect DTAN to classification network.
Minimize classification loss (i.e., cross-entropy)

CNN Classification Experiment

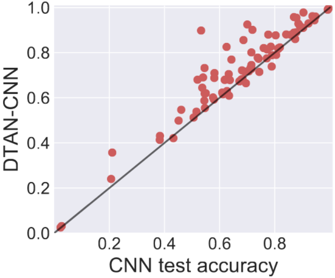


Step 4: use entire model – DTAN-CNN
For classification.



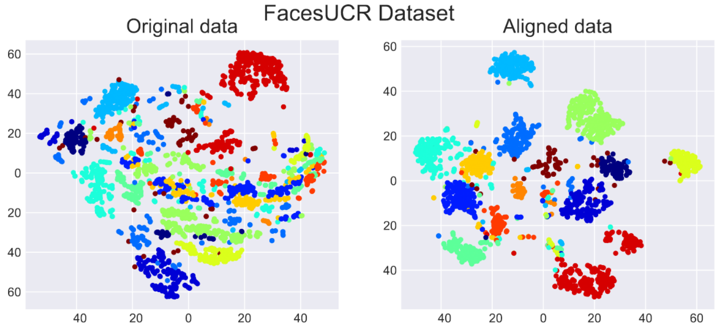
CNN Classification Results

- We compared the average test accuracy of DTAN-CNN to the same CNN without DTAN, using 5 runs per dataset.
- DTAN-CNN achieved higher, or equal to,correct classification rates on 87% of the datasets.



CNN Classification Results

- t-SNE visualization of the original and aligned data, illustrates how DTAN decreases intra-class variance while increasing inter-class one, thus improving the performance of classification net.



Conclusion

- Building on both recent ideas such as STNs, efficient highly-expressive diffeomorphisms and older ones such as congealing we proposed DTAN, a deep net for learning time-series joint alignment.
- The alignment learning is done in an unsupervised way. class labels are used for class partitioning.
- In addition, we proposed a regularization term for the warps, which is critical in an unsupervised framework.
- We also proposed R-DTAN, a recurrent variant of DTAN, which improves the expressiveness and performance of DTAN without increasing the number of parameters.

Thank You

Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

O. Freifeld, S. Hauberg, K. Batmanghelich, and J. W. Fisher III. Highly-expressive spaces of well-behaved transformations: Keeping it simple. In *ICCV*, 2015.

O. Freifeld, S. Hauberg, K. Batmanghelich, and J. W. Fisher III. Transformations based on continuous piecewise-affine velocity fields. *IEEE TPAMI*, 2017.