Dynamical system's modelling for tic activity recognition

Jules Gottraux

EPFL

November 10th, 2020

Introduction

Tic disorders:

- Conditions that induces motor and/or vocal spasms
- Pretty common, especially for young people
- Behavioral therapy as first-line treatment

$\longrightarrow$ Our goal is to facilitate this therapy by automating tic detection

Introduction

Idea from Joseph F. McGuire and Joey Ka-Yee Essoe. Psychologists specialized in neuropsychiatric conditions such as tic disorder in the childhood.

They contribute by creating the dataset for tic detection.

Creation started in september 2020.

Setup for dataset

Previous work

Our baseline is an activity recognition framework on pre-segmented dataset using linear dynamical systems. We'll build our methods from there.

Screenshot of each task in JIGSAWS dataset

Previous work

Fit a linear dynamical system to the video $\boldsymbol{Y} \in \mathbb{R}^{N\times P}$, $\boldsymbol{Y} = \left(y_1, \ldots, y_N\right)^T$:

$$y_t = \boldsymbol{C} x_t$$
$$x_{t+1} = \boldsymbol{A} x_t$$

The projection is obtained by Principal Component Analysis (PCA) and is fixed. The matrix $\boldsymbol{A}$ is then obtained via:

$$\boldsymbol{X}_- \boldsymbol{A} = \boldsymbol{X}_+$$
$$\Rightarrow \boldsymbol{A} = \boldsymbol{X}_-^\dagger \boldsymbol{X}_+$$

$\boldsymbol{X}$ are the encoded frames and $\boldsymbol{X}_- = \left(x_1, \ldots, x_{N-1}\right)^T$, $\boldsymbol{X}_+ = \left(x_2, \ldots, x_N\right)^T$

Outline

Our work consists in improving and adapting this baseline from activity recognition to tic detection and is composed of 4 main sections:

- Baseline extension with non-linear dimensionality reduction
- Baseline extension with joint learning of the dynamical system's components
- Extension of the techniques to online detection
- Test on Hopkins' dataset

All experiments are done on 256×256 videos converted to gray scale.

Non-linear projection

For a video with N frames: $\boldsymbol{Y} \in \mathbb{R}^{N \times P}$. We seek to find a mapping $\boldsymbol{\Phi}_E : \mathbb{R}^P \rightarrow \mathbb{R}^R$, $R \ll P$ such that the frames are *well represented* in the latent space.

We minimize the reconstruction error on a video from Hopkins:

$$\mathscr{L}_{rec} = \frac{1}{N} \sum_{t=1}^{N} \| y_t - \boldsymbol{\Phi}_D(\boldsymbol{\Phi}_E(y_t)) \|_2^2 = \| \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \|_F^2$$

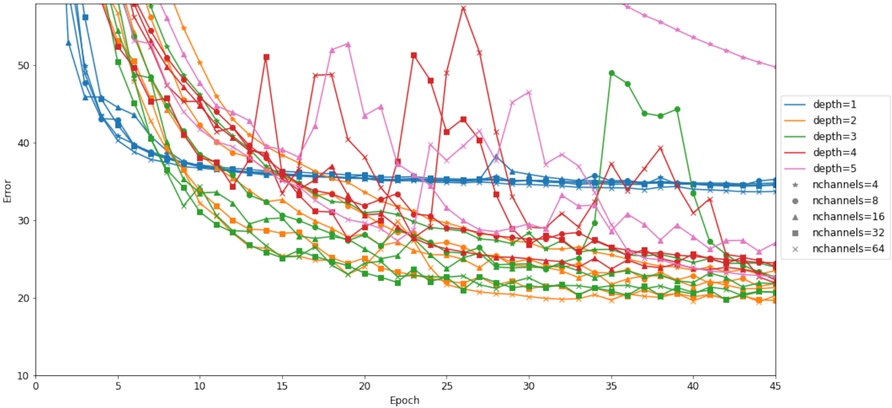
Where $\widehat{\boldsymbol{Y}}$ is the reconstruction of all frames. And $\boldsymbol{\Phi}_D : \mathbb{R}^R \rightarrow \mathbb{R}^P$ is the decoder, i.e. the inverse mapping.

Non-linear projection

We compare three models, each one is a neural network with an autoencoder structure:

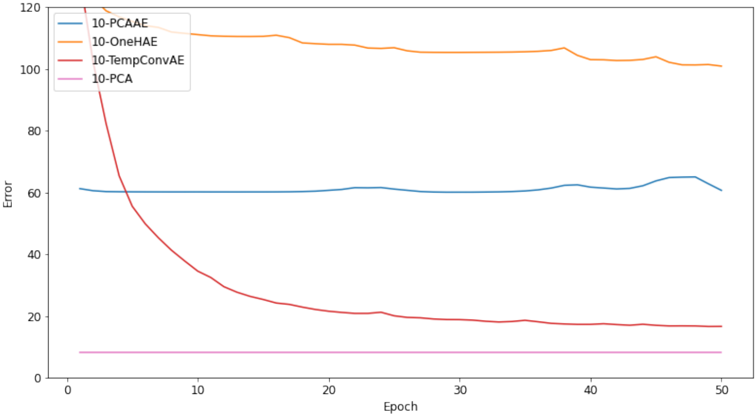
- `PCAAE` (PCA autoencoder): autoencoder using a linear projection
- `OneHAE` (one hidden layer autoencoder): autoencoder using a one hidden layer neural network structure
- `TempConvAE` (temporal convolutional autoencoder): autoencoder using 3d (a.k.a. temporal) convolutional layers

Non-linear projection



Training error for all temporal convolutional networks, $R = 10$

Non-linear projection



Training error for all models, $R = 10$

Non-linear projection

Notes:

- Further analysis showed that enforcing the latent dimension with a linear mapping were hurting the model's power
- Better performance could be obtained using known video compression algorithm or network (e.g. U-Net)
- These preliminary results are not conclusive
- We'll stick to linear projection

Validation and joint learning assessment

Initial goals with this dataset:

- Measure the capability of the method to detect an activity on raw frames
- Assess whether learning $\boldsymbol{A}$ and $\boldsymbol{C}$ jointly helps the algorithm

Instead of fixing the projection and obtaining $\boldsymbol{A}$ as in the baseline, we initialize the models with the parameters from the baseline and minimize:

$$
\begin{aligned}
\mathscr{L}_{pred} &= \frac{1}{N-1}\sum_{t=2}^{N} \| y_t - \boldsymbol{\Phi}_D(\boldsymbol{A}\boldsymbol{\Phi}_E(y_{t-1})) \|_2^2 \\
&= \| \boldsymbol{Y}_+ - \boldsymbol{\Phi}_D(\boldsymbol{\Phi}_E(\boldsymbol{Y}_-)\boldsymbol{A}) \|_F^2
\end{aligned}
$$

$\boldsymbol{\Phi}_E$ and $\boldsymbol{\Phi}_D$ as before are the mapping to and from the latent space.

Validation and joint learning assessment

Evaluation:

- We have fragments of videos, each with a given activity and fitted dynamical system
- Classification task, the features are the dynamical system's matrices and the labels are the corresponding activity

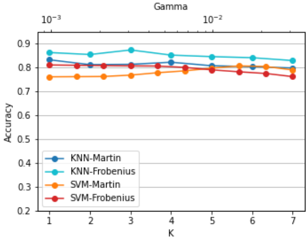
The classification uses metrics based on the subspace angles between dynamical models. These metrics are the Frobenius and Martin distance.

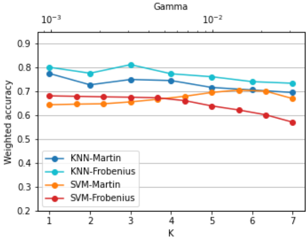
These metrics tells us how much the dynamics of two different fragments differ.

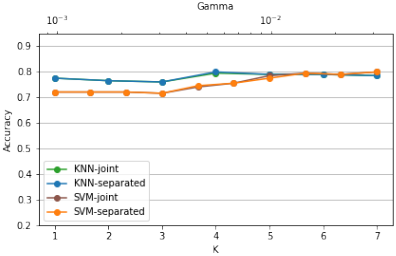
On top of these metrics, we use K-Nearest Neighbors (KNN) or Support Vector Machine with radial basis function kernel (SVM) for the classification

Validation and joint learning assessment

Results for one task (suturing):







Validation and joint learning assessment

Notes:

- Joint learning performance is indistinguishable from separated learning. Hence, separated learning is nearly optimal for this framework
- The model's performance are accurate enough for our detection task

Extension to online detection

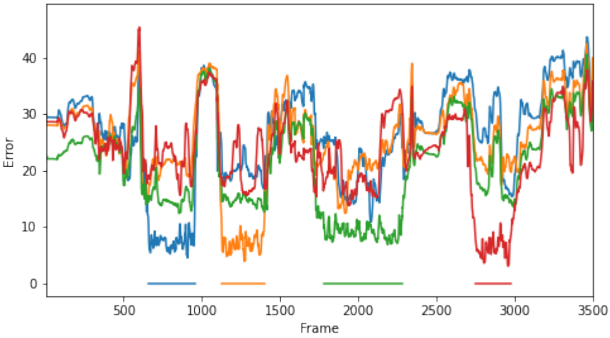
We test two approaches for online detection:

- Detection based on the reconstruction error of the prediction of models
- Detection based on the distance between models and a model fitted in a moving window fashion

Experiments are done on a video picked from the JIGSAWS dataset.

Detection based on reconstruction error

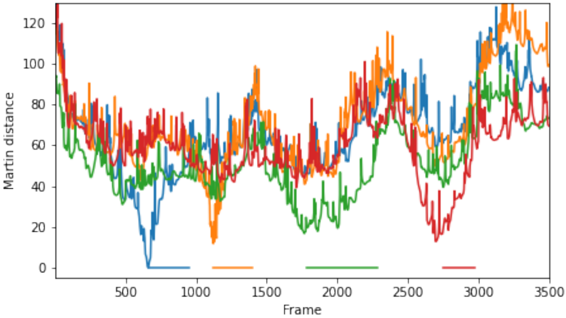
- Pick a particular gesture
- Do the models of this gesture capture the dynamics of the others as well?
- If yes we could use this information to construct an online classifier



Reconstruction error of the prediction of models from same activity

Detection using Martin distance

- Pick a particular gesture
- Can we detect the occurence of this gesture based on the distance between a moving window model and the fitted models?



Martin distance between models of the gesture and moving window model

Hopkins dataset

- 3 videos of 15 minutes each
- Very subtle tics
  - Eye blinking tic
  - Nose flare

Two evaluations:

- Binary classification of tic on pre-segmented fragments
- Comparison of tic occurrence with reconstruction error of normal model

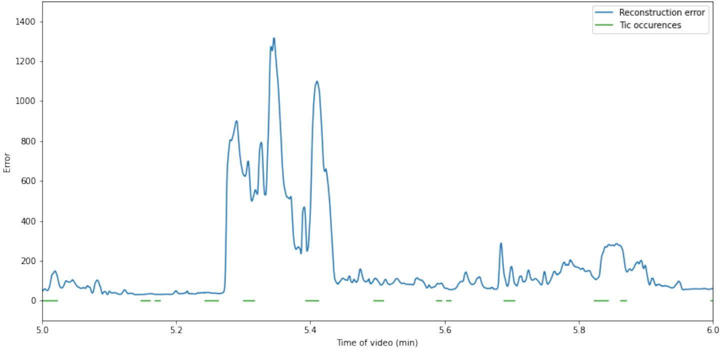
Classification evaluation on Hopkins dataset

Same approach as for JIGSAWS dataset:

- Pre-segmented data
- Using baseline
- KNN or SVM classifier using Martin and Frobenius distance

Evaluation using reconstruction error

- Fit a dynamical system on all *normal* frames (no tic)
- Does the reconstruction error and the occurence of tics correlate?



Reconstruction error of the prediction for normal models

Conclusion

Our experiments indicate that detection of subtle tics is very challenging.

For online detection, creating dynamical models that better captures the general dynamics of the activity is a key ingredient.

Need more data and/or more visible tics to be able to make a proper evaluation.