Supplementary Material for ECCV 2014 Paper: Spatio-temporal Event Classification using Time-series Kernel based Structured Sparsity

László A. Jeni[1], András Lőrincz[2], Zoltán Szabó[3], Jeffrey F. Cohn[1,4], and Takeo Kanade[1]

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary
[3] Gatsby Computational Neuroscience Unit, University College London, London, UK
[4] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
laszlo.jeni@ieee.org, andras.lorincz@elte.hu, zoltan.szabo@gatsby.ucl.ac.uk, jeffcohn@cs.cmu.edu, tk@cs.cmu.edu

Abstract. This document accompanies the paper: "Spatio-temporal Event Classification using Time-series Kernel based Structured Sparsity". We provide implementation details of the Kernel Structured Sparsity (KSS) method for FISTA (fast iterative shrinkage-thresholding algorithm) [2]. We also show scaling results on the 6D Motion Gesture Database [3]. Note that information given in this document is not necessary to understand the content of the main paper.

S1 Implementation

We can rewrite our optimization cost function (Eq. (S1)) as

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\left\langle \varphi(x) - \sum_{m=1}^{M}\varphi(d_i)\alpha_i, \varphi(x) - \sum_{i=1}^{M}\varphi(d_i)\alpha_i \right\rangle_{\mathcal{H}} + \kappa\Omega(\boldsymbol{\alpha}) \tag{S1}$$

$$= \frac{1}{2}\Bigg[\langle \varphi(x), \varphi(x)\rangle_{\mathcal{H}} - 2\left\langle \varphi(x), \sum_{i=1}^{M}\varphi(d_i)\alpha_i \right\rangle_{\mathcal{H}} + \left\langle \sum_{i=1}^{M}\varphi(d_i)\alpha_i, \sum_{i=1}^{M}\varphi(d_i)\alpha_i \right\rangle_{\mathcal{H}} \Bigg] + \kappa\Omega(\boldsymbol{\alpha})$$

$$= \frac{1}{2}\Bigg[\langle \varphi(x), \varphi(x)\rangle_{\mathcal{H}} - 2\sum_{i=1}^{M} k(x, d_i)\alpha_i + \sum_{i=1}^{M}\sum_{j=1}^{M}\alpha_i k(d_i, d_j)\alpha_j \Bigg] + \kappa\Omega(\boldsymbol{\alpha})$$

$$= \frac{1}{2}\left[-2\mathbf{k}^T\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\alpha}\right] + \kappa\Omega(\boldsymbol{\alpha}) + c, \tag{S2}$$

where we used that the kernel represents an inner product and bilinearity of the inner product was exploited, superscript 'T' denotes transposition, we introduced the

$$\mathbf{k} = [k(\mathbf{x}, \mathbf{d}_1); \ldots; k(\mathbf{x}, \mathbf{d}_M)] \in \mathbb{R}^M, \tag{S3}$$

$$\mathbf{G} = [G_{ij}] = [k(\mathbf{d}_i, \mathbf{d}_j)] \in \mathbb{R}^{M \times M} \tag{S4}$$

notations, and c is an additive constant independent of $\boldsymbol{\alpha}$. By discarding c, our objective function is of the form

$$J(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}) + \kappa \Omega(\boldsymbol{\alpha}), \tag{S5}$$

where

$$f(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - \mathbf{k}^T \boldsymbol{\alpha} \tag{S6}$$

is a quadratic function.

Using this form, a FISTA optimization can be adapted to the solution. Our experiments were based on the modification of the SLEP package[5]. We need the following elements for the implementation:

1. The proximal operator of Ω (it has not changed).
2. $f(\boldsymbol{\alpha})$ from Eq. (S6)
3. The gradient of f:

$$\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) = \mathbf{G}\boldsymbol{\alpha} - \mathbf{k}. \tag{S7}$$

4. The stopping criterion for FISTA.

Furthermore, for practical issues, it is important to identify a set of candidate values for the regularization parameter κ, since it affects the performance of the fitted model. We address the stopping criterion and the regularization parameter selection in the following two subsections.

S1.1 FISTA Stopping Criterion

The principle behind FISTA is to iteratively form quadratic approximations $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to $J(\boldsymbol{\alpha})$ around a carefully chosen point $\boldsymbol{\beta}$, and to minimize $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ rather than the original cost function $J(\boldsymbol{\alpha})$. We can stop the iterations if the relative change in $\boldsymbol{\alpha}$ between consecutive iterations being sufficiently small.

Lets define $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as:

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \langle \boldsymbol{\alpha} - \boldsymbol{\beta}, \nabla f(\boldsymbol{\beta}) \rangle + \frac{L}{2} \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|^2 + \kappa \Omega(\boldsymbol{\alpha}). \tag{S8}$$

where $L = L(f)$ is a Lipschitz constant $L(f)$.

To define the stopping criterion we need the following fundamental property for a smooth function [2]:

[5] http://www.public.asu.edu/~jye02/Software/SLEP/

Lemma 1. *Let $f(\boldsymbol{\alpha}) : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with Lipschitz continuous gradient and Lipschitz constant $L(f)$. Then, for any $L \geq L(f)$,*

$$f(\boldsymbol{\alpha}) \leq f(\boldsymbol{\beta}) + \langle \boldsymbol{\alpha} - \boldsymbol{\beta}, \nabla f(\boldsymbol{\beta}) \rangle + \frac{L}{2} \left\| \boldsymbol{\alpha} - \boldsymbol{\beta} \right\|^2, \tag{S9}$$

for every $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$

Applying Lemma 1 to (Eq. (S6)) we get

$$\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - \mathbf{k}^T \boldsymbol{\alpha} \leq \frac{1}{2}\boldsymbol{\beta}^T \mathbf{G} \boldsymbol{\beta} - \mathbf{k}^T \boldsymbol{\beta} + (\boldsymbol{\alpha} - \boldsymbol{\beta})^T (\mathbf{G}\boldsymbol{\beta} - \mathbf{k}) + \frac{L}{2} \left\| \boldsymbol{\alpha} - \boldsymbol{\beta} \right\|^2, \tag{S10}$$

$$\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - \mathbf{k}^T \boldsymbol{\alpha} \leq \frac{1}{2}\boldsymbol{\beta}^T \mathbf{G} \boldsymbol{\beta} - \mathbf{k}^T \boldsymbol{\beta} + \tag{S11}$$

$$\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\beta} - \boldsymbol{\alpha}^T \mathbf{k} - \boldsymbol{\beta}^T \mathbf{G} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{k} + \frac{L}{2} \left\| \boldsymbol{\alpha} - \boldsymbol{\beta} \right\|^2, \tag{S12}$$

$$\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} \leq \boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^T \mathbf{G} \boldsymbol{\beta} + \frac{L}{2} \left\| \boldsymbol{\alpha} - \boldsymbol{\beta} \right\|^2, \tag{S13}$$

$$\tag{S14}$$

which leads to the stopping criterion:

$$\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{G} \boldsymbol{\beta} \leq L \left\| \boldsymbol{\alpha} - \boldsymbol{\beta} \right\|^2. \tag{S15}$$

S1.2 Candidate Values for Regularization Parameter $\boldsymbol{\kappa}$

It is important to choose an appropriate regularization parameter from a set of candidate values, because it affects the predictive performance of the fitted model. Since the derived objection function (Eq. (13)) is of the form of Eq. (1.1) in [1], thus in case of

$$\Omega^*(\nabla f(\mathbf{0})) \leq \kappa \tag{S16}$$

the solution is guaranteed to be $\boldsymbol{\alpha} = \mathbf{0}$ (see the notes for Proposition 1.2 in [1]).

Let Ω be the l_1/l_q norm ($q \geq 1$) induced by a $\mathcal{G} = \{\{1, \ldots, d\}, \{d+1, \ldots, 2d\}, \ldots, \{(M-1)d+1, \ldots, Md\}\}$ partition

$$\Omega(\mathbf{u}) = \sum_{G \in \mathcal{G}} \left\| \mathbf{u}_G \right\|_q. \tag{S17}$$

Then

$$\Omega^*(\mathbf{u}) = \max_{G \in \mathcal{G}} \left\| \mathbf{u}_G \right\|_{q'} \tag{S18}$$

is the l_∞/l'_q norm, where q are q' conjugate, i.e.,

$$\frac{1}{q} + \frac{1}{q'} = 1. \tag{S19}$$

For example, in case of $q = 2$, $q' = 2$ and using $\nabla f(\mathbf{0}) = -\mathbf{k}$ (Eq. (S7)), and for

$$\max_{G \in \mathcal{G}} \|\mathbf{k}_G\|_2 \leq \kappa \tag{S20}$$

$\boldsymbol{\alpha} = \mathbf{0}$ is guaranteed. In other words, the application of a

$$\kappa = \kappa' \max_{G \in \mathcal{G}} \|\mathbf{k}_G\|_2 \tag{S21}$$

parameterization is advisable, where $\kappa' \in (0, 1)$.

S2 Scalability Experiment on the 6DMG

In this set of experiment we studied the scalability of the proposed methods by varying the dimensions of the optimization problems. We measured the performances of the methods for gesture classification. We calculated Gram matrices using the GA kernel from the time-series provided with the dataset and performed leave-one-subject out cross validation. We searched for the best parameter (σ of GA kernel) between 0.4 and 20 and selected the parameter having the lowest mean classification error. The SVM regularization parameter (C) was searched within 2^{-10} and 2^{10} and the KSS regularization parameter (κ) was searched within 0 and 0.5 in a similar fashion. We summarize the results in Table S1.

Table S1. The character error rates (CER) of motion character recognition using single modalities with varying numbers of classes. The different attributes in the columns: position (P), velocity (V), acceleration (A), angular velocity (W), orientation (O). The best results are denoted with bold letters.

# of Classes	Classifier	*P*	*W*	*O*	*A*	*V*
3	GA + SVM	0.53	**0.27**	0.8	0.537	**0.27**
	GA + KSS	**0.13**	**0.13**	0.53	**0.13**	0.27
10	GA + SVM	2.08	**1.8**	4.52	2.32	**1.8**
	GA + KSS	0.8	**0.56**	4.04	0.96	0.84
26	GA + SVM	**3.68**	4.83	7.82	6.15	3.69
	GA + KSS	**3.43**	4.95	13.15	4.8	3.88

In all experiments the KSS method outperformed the time-series SVM. In the cases with 3 and 10 class labels, the KSS method achieved a two- and three-fold reduction in recognition error (respectively).

Regarding the parameters, we note that the KSS achieved the best results with $\kappa' \geq 0$ values, which supports the relevancy of the structured sparse regularization.

Figure S1 shows the confusion matrix of the KSS method using all the 26 class labels. Most gestures were predicted with 100% accuracy, while some of the similar looking gesture-pairs, such as D-P, K-R, U-V and Q-G, show some misclassification error.

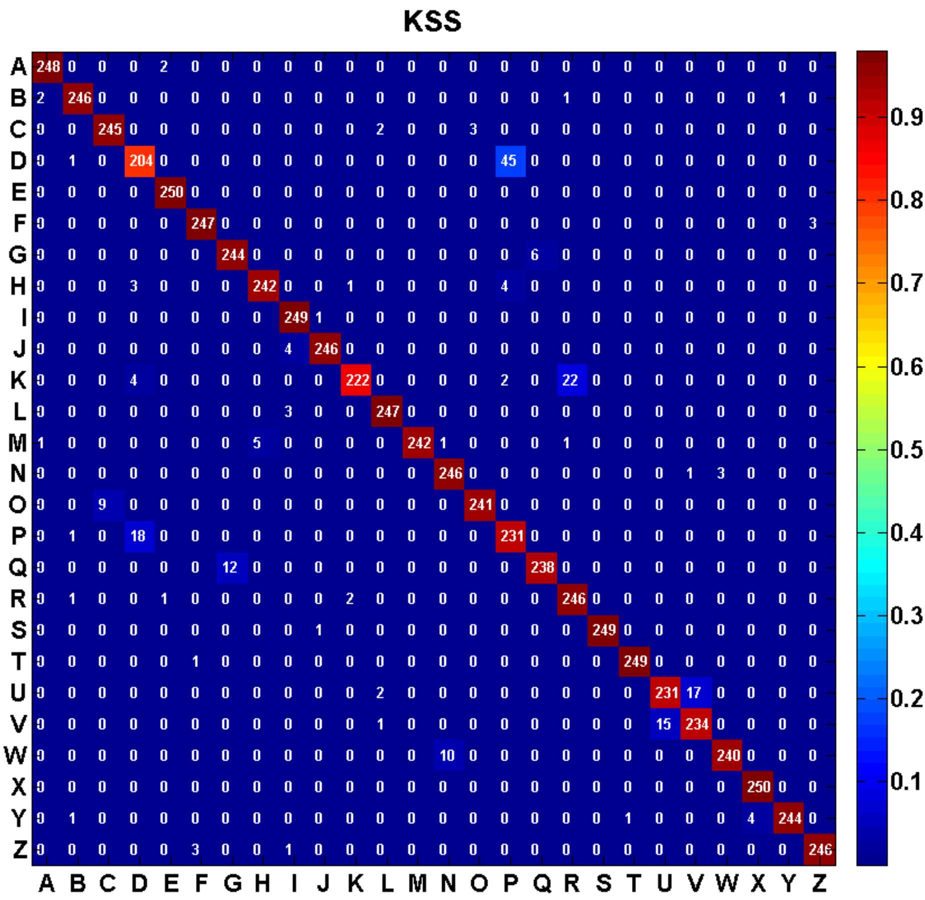

Fig. S1. Confusion matrix for gesture classification on the 6DMG dataset using Kernel Structured Sparsity (KSS) applied to position data (P).

References

1. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning 4, 1–106 (2012)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2, 183–202 (2009)
3. Chen, M., AlRegib, G., Juang, B.H.: 6dmg: A new 6d motion gesture database. In: 3rd Multimedia Systems Conference. pp. 83–88. MMSys '12, ACM, New York, NY, USA (2012)