Spatio-temporal Event Classification using Time-series Kernel based Structured Sparsity

László A. Jeni — Carnegie Mellon University — *laszlo.jeni@ieee.org*

András Lőrincz — Eötvös Loránd University — *andras.lorincz@elte.hu*

Zoltán Szabó — University College London — *zoltan.szabo@gatsby.ucl.ac.uk*

Jeffrey F. Cohn — University of Pittsburgh — *jeffcohn@cs.cmu.edu*

Takeo Kanade — Carnegie Mellon University — *tk@cs.cmu.edu*

Code is available online at https://github.com/laszlojeni/KSS

1. Introduction

In many behavioral domains, such as facial expression and gesture, sparse structure is prevalent. This sparsity would be well suited for event detection but for one problem. Features typically are confounded by alignment error in space and time. As a consequence, high-dimensional representations such as SIFT and Gabor features have been favored despite their much greater computational cost and potential loss of information. We propose a Kernel Structured Sparsity (KSS) method that can handle both the temporal alignment problem and the structured sparse reconstruction within a common framework, and it can rely on simple features.

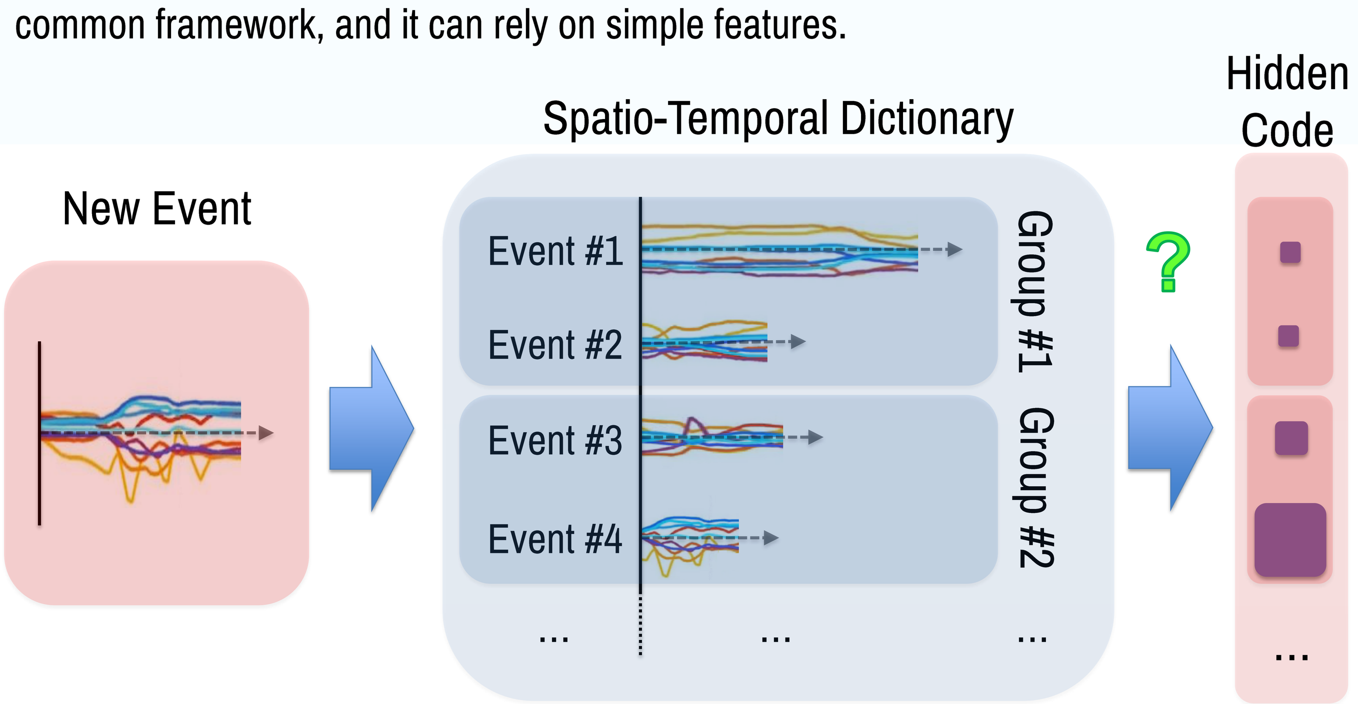


Fig.1. The goal is to approximate spatio-temporal events by a few groups of such events.

2. Face Alignment

We used Zface (www.zface.org), which is a generic 3D face tracker that requires no individual training to track facial landmarks of persons is has never seen before. It locates 3D coordinates of a dense set of facial landmarks.

The 3D point distribution model (PDM) describes non-rigid shape variations linearly and composes it with a global rigid transformation, placing the shape in the image frame:

$$\mathbf{x}_i = \mathbf{x}_i(\mathbf{p}) = s\mathbf{R}(\bar{\mathbf{x}}_i + \boldsymbol{\Phi}_i\mathbf{q}) + \mathbf{t} \quad (i = 1, \ldots, M),$$

where $x_i(\mathbf{p})$ denotes the 3D location of the i^{th} landmark and $p = \{s, \alpha, \beta, \gamma, \mathbf{q}, \mathbf{t}\}$ denotes the parameters of the model, which consist of a global scaling s, angles of rotation in three dimensions, a translation $\mathbf{t}$ and non-rigid transformation $\mathbf{q}$.

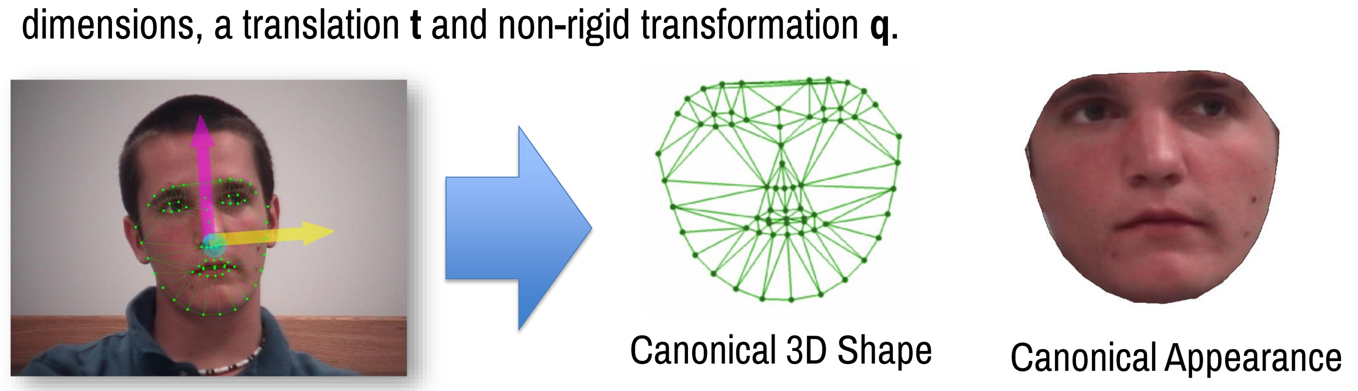


Fig.2. 3D face alignment and canonical views.

3. Time-series Building

We tracked the video sequences with the ZFace tracker and built the time-series from the PCA coefficients of the 3D PDM (parameter $\mathbf{q}$). Illustratively, this is the compressed representation of the 3D landmark locations without rigid head movements.

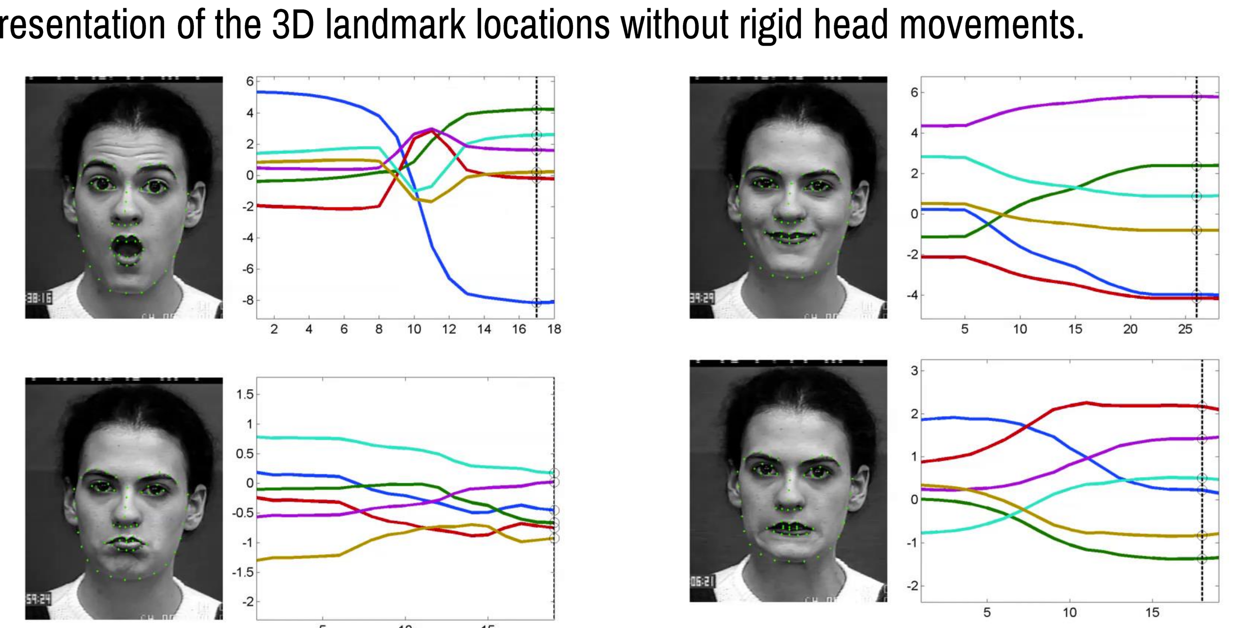
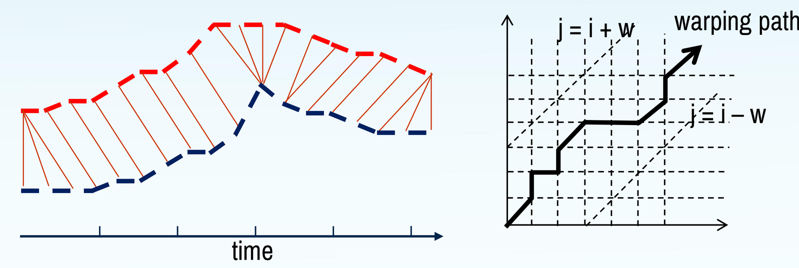


Fig.3. Holistic facial expressions from CK+ and the corresponding time-series.

4. Global Alignment Kernel

To quantify the similarity of time-series (that form the input of the classifiers) we make use of kernels. Let $|\pi|$ denote the length of alignment π. The cost can be defined by means of a local divergence ϕ that measures the discrepancy between any two points u_i and v_j of vectors $\mathbf{u}$ and $\mathbf{v}$.



The Global Alignment (GA) kernel assumes that the minimum value of alignments may be sensitive to peculiarities of the time series and intends to take advantage of all alignments weighted exponentially. It is defined as the sum of exponentiated and sign changed costs of the individual alignments.

$$k_{GA}(\mathbf{u}, \mathbf{v}) = \sum_{\pi \in A(n,m)} \prod_{i=1}^{|\pi|} e^{-\phi(u_{\pi_1(i)}, v_{\pi_2(i)})}$$

$$\phi_\sigma(x, y) = \frac{1}{2\sigma^2}\|x - y\|^2 + \log\left(2 - e^{-\frac{\|x-y\|^2}{2\sigma^2}}\right)$$

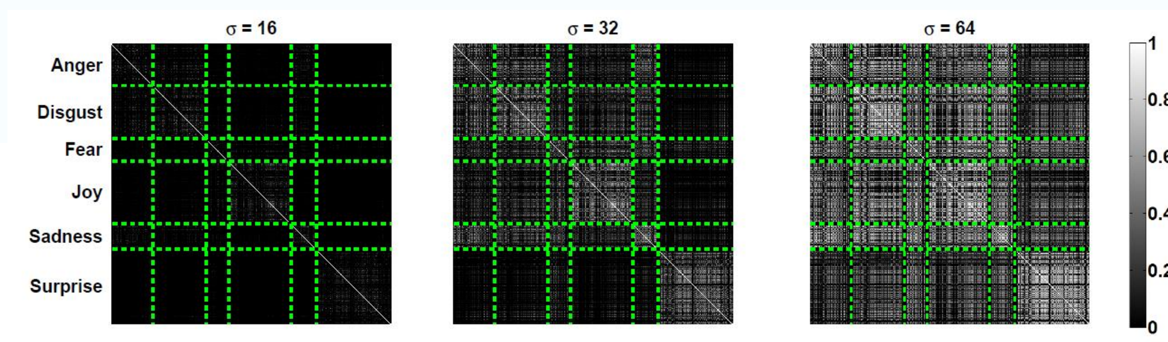


Fig.4. Gram matrices induced by the GA kernel with different parameters.

5. Kernel Structured Sparsity

Euclidean spaces

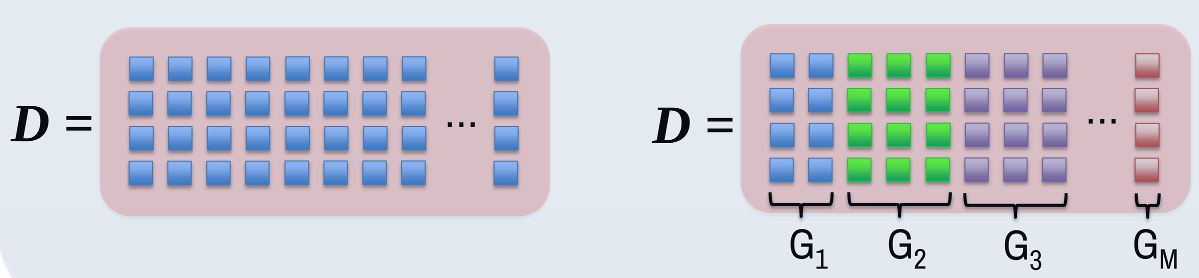
Lasso (R. Tibshirani.,1996)

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\|x - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \to \min_{\boldsymbol{\alpha}}$$

Group Lasso (M. Yuan & Y. Lin, 2006)

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\|x - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \kappa\Omega(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha}}$$

$$\Omega(\boldsymbol{\alpha}) = \left\|\left(\|\boldsymbol{\alpha}_G\|\right)_{G\in\mathcal{G}}\right\|_q \quad (q \geq 1).$$



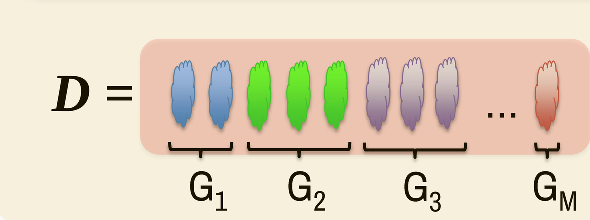
Hilbert space

Kernel Structured Sparsity

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\left\|\varphi(x) - \sum_{i=1}^{M}\varphi(d_i)\alpha_i\right\|_{\mathcal{F}}^2 + \kappa\Omega(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha}}$$

$$\varphi : (\mathbb{R}^d)^{\mathbb{N}} \to \mathcal{H}$$

$$\Omega(\boldsymbol{\alpha}) = \left\|\left(\|\boldsymbol{\alpha}_G\|\right)_{G\in\mathcal{G}}\right\|_q \quad (q \geq 1).$$



Kernel trick

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

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

Quadratic function!
This is a finite dimensional problem.

Notations:

$$\mathbf{k} = [k(\mathbf{x}, \mathbf{d}_1); \ldots; k(\mathbf{x}, \mathbf{d}_M)] \in \mathbb{R}^M,$$
$$\mathbf{G} = [G_{ij}] = [k(\mathbf{d}_i, \mathbf{d}_j)] \in \mathbb{R}^{M\times M}$$

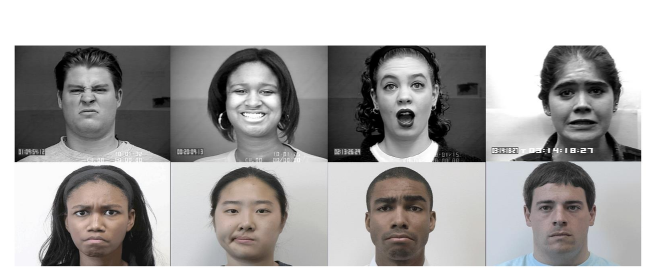
Using this form, a **FISTA** (fast iterative shrinkage-thresholding algorithm) optimization can be adapted to the solution. Our experiments were based on the modification of the SLEP package.

We need the following elements for the implementation:

1. The proximal operator of Ω (it has not changed)
2. $f(\alpha)$ from the cost function
3. The gradient of f: $\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) = \mathbf{G}\boldsymbol{\alpha} - \mathbf{k}$.
4. The stopping criterion for FISTA (see supplement).

6. Datasets

Cohn-Kanade Extended (holistic facial expressions)



Emotion	N
Angry (An)	45
Contempt (Co)	18
Disgust (Di)	59
Fear (Fe)	25
Happy (Ha)	69
Sadness (Sa)	28
Surprise (Su)	83

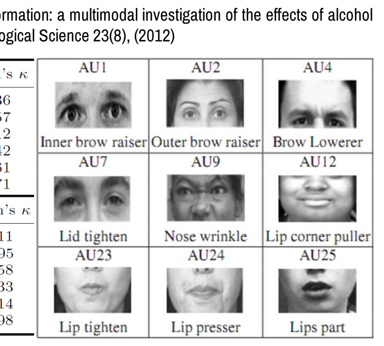
* P. Lucey et al.: The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression

Group Formation Task (action units)

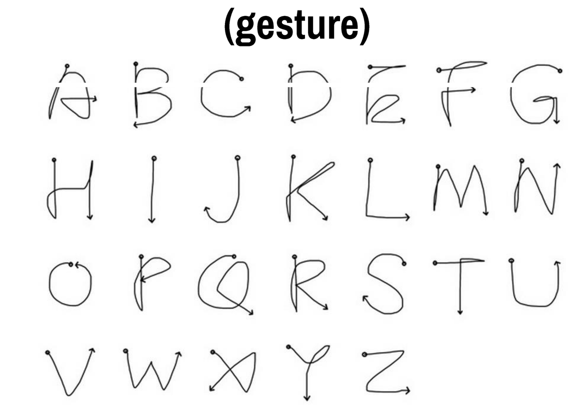
* M. Sayette et al.: Alcohol and group formation: a multimodal investigation of the effects of alcohol on emotion and social bonding. Psychological Science 23(8), (2012)

AU	FACS Name	Cohen's κ
1	Inner Brow Raiser	0.936
2	Outer Brow Raiser	0.857
4	Brow Lowerer	0.912
7	Lid Tightener	0.942
10	Upper Lip Raiser	0.961
11	Nasolabial Deepener	0.971

AU	FACS Name	Cohen's κ
12	Lip Corner Puller	0.911
14	Dimpler	0 895
16	Lower Lip Depressor	0.858
17	Chin Raiser	0.833
20	Lip Stretcher	0.914
22	Lip Funneler	0.798



6DMG Air-handwriting (gesture)



*M. Chen, G. AlRegib and B-H. Juang: 6DMG: A New 6D Motion Gesture Database (MMSYS12)

Database	Domain	Type	# of Time-series	# of Classes	Dimension	Avg. length (std)
6DGM [9]	Gesture	Deliberate	5720	26	3-4	66.86 (29.84)
CK+ [24]	Face	Deliberate	327	7	56	17.97 (8.59)
GFT50 [32]	Face	Spontaneous	5000	12	166	7.51 (1.44)

7. Results

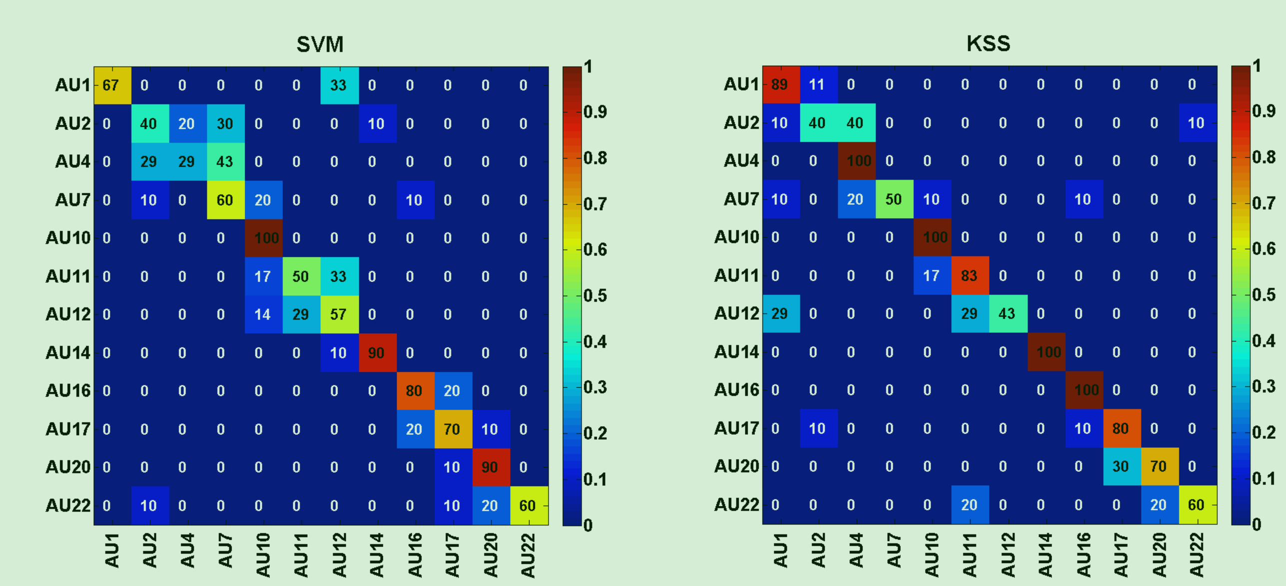
Cohn-Kanade Extended

	Non-sparse											Sparse		
	Frame level							Fixed length			TS	Frame level		TS
	3D Shape [17]	Gabor [33]	LBP [33]	MSDF [33]	Simple BoW [33]	SS-SIFT+BoW [33]	MSDF+BoW [33]	Gabor [38]	ICA [21]	Dynamic Haar [40]	**3D Shape + GA (this work)**	Gabor [25]	Shape [42]	**3D Shape + GA + KSS (this work)**
Acc.	86.8	91.81	82.38	94.34	92.67	93.28	95.85	-	-	-	**97.9**	93.8	92.4	**97.6**
AUC	-	-	-	-	-	-	-	.978	.978	.966	**.991**	-	-	-

Metric	SVM	KSS-1	KSS-2	KSS-3
Macro F_1	**0.909**	0.881	0.889	0.902
Micro F_1	**0.935**	0.916	0.922	0.932
Avg. TPR	**0.900**	0.868	0.877	0.896

Metric	SVM	KSS-1	KSS-2	KSS-3
Macro F_1	0.658	**0.743**	0.653	0.664
Micro F_1	0.679	**0.761**	0.679	0.688
Avg. TPR	0.660	**0.763**	0.661	0.669

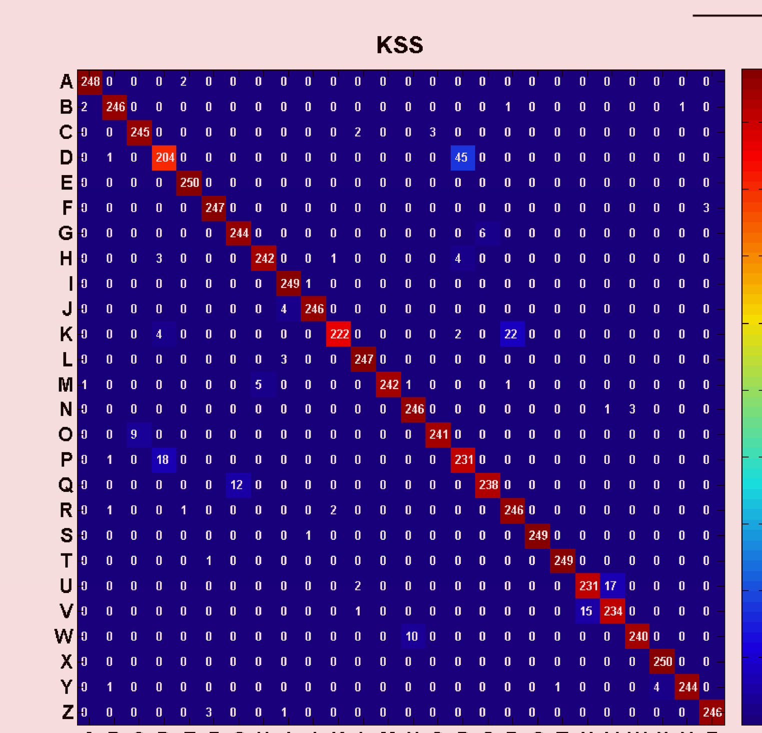
Group Formation Task (action units)



6DMG Air-handwriting (gesture)

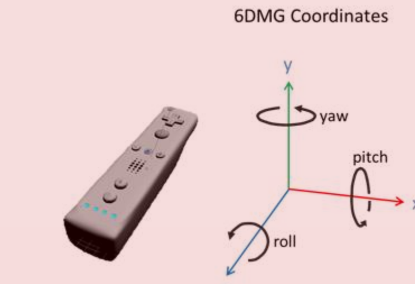
Classifier	P	W	O	A	V
Chen [8] (HMM)	3.72	7.92	**3.81**	7.97	6.12
This work (SVM)	3.68	**4.83**	7.82	6.15	**3.69**
This work (KSS)	**3.43**	4.95	13.15	**4.8**	3.88

# 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



Features:

- P: position in 3D
- W: angular velocity
- O: absolute orientation
- A: acceleration
- V: velocity



Acknowledgments: Research reported in this publication was supported in part by the National Institute of Mental Health of the National Institutes of Health under Award Number MH096951; the Research and Technology Innovation Fund (grant agreement EITKIC_12-1-2012-0001), the EIT ICT Labs; and the Gatsby Charitable Foundation.