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,tk}@cs.cmu.edu

Abstract. 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. We characterize spatio-temporal events as time-series of motion patterns and by utilizing time-series kernels we apply standard structured-sparse coding techniques to tackle this important problem. We evaluated the KSS method using both gesture and facial expression datasets that include spontaneous behavior and differ in degree of difficulty and type of ground truth coding. KSS outperformed both sparse and non-sparse methods that utilize complex image features and their temporal extensions. In the case of early facial event classification KSS had 10% higher accuracy as measured by F_1 score over kernel SVM methods[1].

Keywords: structured sparsity, time-series kernels, facial expression classification, gesture recognition.

1 Introduction

The analysis and identification of spatio-temporal processes are of great importance in facial expression identification. The change of pixel intensities around 3D landmark points of the face, such as the corners of the mouth or eyes or the motion patterns of the 3D landmark points themselves, are the natural descriptors of the phenomena. The problem is quite challenging, since individual patch

* Electronic supplementary material -Supplementary material is available in the online version of this chapter at http://dx.doi.org/10.1007/978-3-319-10593-2_10. Videos can also be accessed at http://www.springerimages.com/videos/978-3-319-10592-5

[1] The KSS code is available online at https://github.com/laszlojeni/KSS

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 135–150, 2014.

series or temporal series of 3D meshes are to be compared. A further sophistication appears by the changes of pace of any expression. Consider winking for example. It may be longer or shorter, and within a broad range of duration it can have identical (social) meaning. In turn, we have to generalize the recognition procedure over temporally warped signals.

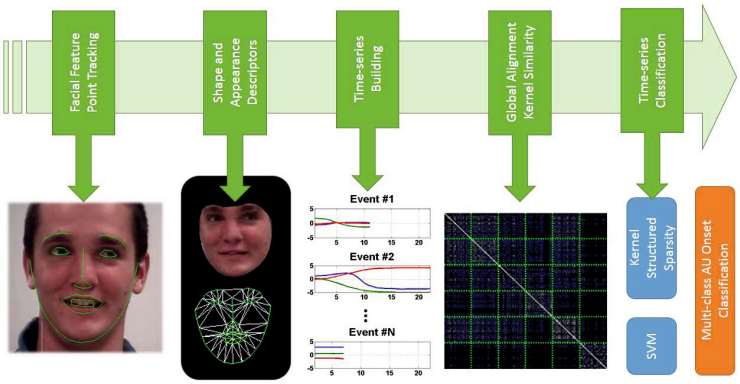

Fig. 1. Overview of the system

Efficient methods using independent component analysis [21], Haar filters [40], and hidden Markov models [4,5,37] have been applied to problems related to the estimation of facial expressions. For temporal clustering of human motion data hierarchical cluster analysis showed promising results [13]. Dynamic time warping is one of the most efficient methods that offer the comparison of temporally distorted samples [30]. Recently developed robust versions, such as the dynamic time warping (DTW) kernel and the global alignment (GA) kernel [10,11] have a great promise here.

In a recent work [22], the authors studied both DTW and GA kernels together with support vector machines (SVM) for holistic facial expressions. Performance was excellent with slight advantage for the GA kernel.

Our interest is in general social conversations, and thus we are interested in the recognition of facial actions (cf. Facial Action Units (AU) [29,28]). The dynamics in this space can reveal emotion, pain, and cognitive states in individuals and the quality of interaction between them. AUs form a large space. There are 30 or so AUs and they can combine in non-additive ways. The recognition problem is demanding. We have studied and compared two methods: (1) the multi-class Support Vector Machine procedure, where for n-classes $n(n-1)/2$ classifiers are developed and evaluation is followed by a voting procedure. This method scales quadratically with the number of classes and slows down considerably as n grows. The other method we studied (2) is structured sparse representation, where the different classes compete with each other and this competition enhances the contrast between the groups. The contrast enhancement is then followed by the winner selection step. We note that this method is also attractive since it easily generalizes to the multi-label situation.

Sparse coding [36,6] and it's structured sparse extensions [41,2] are actively researched topics in machine learning. In the original formulation, we approximate the observations with a linear combination of atoms selected from a fixed dictionary [35], whereas in the case of structured sparsity, the hidden code (i.e., the representation) can be organized into disjoint groups or even into hierarchical structures, such as trees. Furthermore, sparse recovery may involve large ensembles of kernel machines [19]. It was shown recently [2,15,27] that structured sparsity allows significant performance gains beyond the simplistic sparse and compressible models that dominate the literature. For a more detailed survey see [1]. We will use a particular and convenient form. We will optimize structured sparse recovery in a feature space defined by the time-series kernel.

Our contributions are as follows:

Time-series based analysis. Previous work on the use of sparse representation for facial expression recognition [42,25] was limited to individual frames without respect to temporal organization. Our goal is analysis of time series.

Implicit reconstruction in the time-series space. By applying time-series kernels, we can implicitly take into account spatio-temporal similarities. According to our extensive numerical experiments, this method is advantageous in the studied applications.

Structured sparse coding. We show that structured sparse coding is competitive with multi-class SVM method for holistic expressions, facial action units, and also hand gestures. For holistic expressions performance was comparable. For AUs the structured sparse method was 10% more accurate than multi-class SVM. For hand gestures KSS was better with a slight margin.

The paper is organized as follows (see Fig. 1 for a high-level summary of the proposed estimation): Time series of landmark points are used to represent the evolution of facial expressions; this is the topic of Section 2.1. To measure the similarity of the time-series representations we apply global alignment kernels (Section 2.2). Support vector machines and the proposed structured-sparse (KSS) coding technique based on time-series kernels are detailed in Section 2.3 and Section 2.4, respectively. The efficiency of our novel solution method is illustrated by numerical experiments in numerous spatio-temporal gesture and facial expression classification problems in Section 3. Conclusions are drawn in Section 4.

Notations. Vectors ($\mathbf{a}$) and matrices ($\mathbf{A}$) are denoted by bold letters. An $\mathbf{u} \in \mathbb{R}^d$ vector's Euclidean norm is $\|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^{d} u_i^2}$ and ℓ_q norm is $\|\mathbf{u}\|_q = \left(\sum_{i=1}^{d} |u_i|^q\right)^{\frac{1}{q}}$ ($q \geq 1$). Vector $\mathbf{u}_G$ is the restriction of $\mathbf{u}$ to $G \subseteq \{1, \ldots, d\}$. $\mathbf{B} = [\mathbf{A}_1; \ldots; \mathbf{A}_K] \in \mathbb{R}^{(d_1+\ldots+d_K)\times N}$ denotes the concatenation of matrices $\mathbf{A}_k \in \mathbb{R}^{d_k \times N}$. The transpose of vector $\mathbf{u} \in \mathbb{R}^d$ is $\mathbf{u}^T$.

2 Methods

In this section we detail the components of our proposed approach.

2.1 Facial Feature Point Localization

We use representations based on facial feature points, landmarks to describe the evolution of facial events.

To localize a dense set of facial landmarks, Active Appearance Models (AAM) [26], Constrained Local Models (CLM) [31] and Supervised Descent Methods (SDM) [39] are often used. These methods register a dense parameterized shape model to an image such that its landmarks correspond to consistent locations on the face.

Of the two, person specific AAMs have higher precision than CLMs or SDMs, but they must be trained for each person before use. On the other hand, CLM and SDM methods can be used for person-independent face alignment because of the localized region templates.

In this work we used a combined 3D SDM method, where the shape model is defined by a 3D mesh and, in particular, by the 3D vertex locations of the mesh, called landmark points. Consider the shape of a 3D SDM as the coordinates of 3D vertices that make up the mesh:

$$\mathbf{x} = [x_1; y_1; z_1; \ldots; x_M; y_M; z_M], \tag{1}$$

or, $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_M]$, where $\mathbf{x}_i = [x_i; y_i; z_i]$. We have T samples: $\{\mathbf{x}(t)\}_{t=1}^{T}$. We assume that – apart from scale, rotation, and translation – all samples $\{\mathbf{x}(t)\}_{t=1}^{T}$ can be approximated by means of the linear principal component analysis (PCA).

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:

$$\mathbf{x}_i = \mathbf{x}_i(\mathbf{p}) = s\mathbf{R}(\bar{\mathbf{x}}_i + \boldsymbol{\Phi}_i\mathbf{q}) + \mathbf{t} \quad (i = 1, \ldots, M), \tag{2}$$

where $\mathbf{x}_i(\mathbf{p})$ denotes the 3D location of the i^{th} landmark and $\mathbf{p} = \{s, \alpha, \beta, \gamma, \mathbf{q}, \mathbf{t}\}$ denotes the parameters of the model, which consist of a global scaling s, angles of rotation in three dimensions ($\mathbf{R} = \mathbf{R}_1(\alpha)\mathbf{R}_2(\beta)\mathbf{R}_3(\gamma)$), a translation $\mathbf{t}$ and non-rigid transformation $\mathbf{q}$. Here $\bar{\mathbf{x}}_i$ denotes the mean location of the i^{th} landmark (i.e. $\bar{\mathbf{x}}_i = [\bar{x}_i; \bar{y}_i; \bar{z}_i]$ and $\bar{\mathbf{x}} = [\bar{\mathbf{x}}_1; \ldots; \bar{\mathbf{x}}_M]$).

We assume that the prior of the parameters follow a normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{\Lambda}$ at a parameter vector $\mathbf{q}$: $p(\mathbf{p}) \propto N(\mathbf{q}; \mathbf{0}, \boldsymbol{\Lambda})$ and we used PCA to determine the d pieces of $3M$ dimensional basis vectors ($\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1; \ldots; \boldsymbol{\Phi}_M] \in \mathbb{R}^{3M \times d}$). Vector $\mathbf{q}$ represents the 3D distortion of the face in the $3M \times d$ dimensional subspace and it can be used for emotion classification, for example.

We used ZFace[2], 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. We note that the 3D PDM of ZFace is consists of 56 non-rigid parameters ($\mathbf{q} \in \mathbb{R}^{56}$).

[2] ZFace is available from `http://zface.org`.

2.2 Global Alignment Kernel

To quantify the similarity of time-series (that form the input of the classifiers) we make use of kernels.

Kernel based classifiers, like any other classification scheme, should be robust against invariances and distortions. Dynamic time warping, traditionally solved by dynamic programming, has been introduced to overcome temporal distortions and has been successfully combined with kernel methods.

Let $\mathcal{X}^{\mathbb{N}}$ be the set of discrete-time time series taking values in an arbitrary space $\mathcal{X}$. One can try to align two time series $\mathbf{u} = (u_1, ..., u_n)$ and $\mathbf{v} = (v_1, ..., v_m)$ of lengths n and m, respectively, in various ways by distorting them. An alignment π has length p and $p \leq n + m - 1$ since the two series have $n + m$ points and they are matched at least at one point of time. We use the notation of [12]. An alignment π is a pair of increasing integral vectors (π_1, π_2) of length p such that $1 = \pi_1(1) \leq ... \leq \pi_1(p) = n$ and $1 = \pi_2(1) \leq ... \leq \pi_2(p) = m$, with unitary increments and no simultaneous repetitions. Coordinates of π are also known as warping functions.

Now, 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 $\mathbf{v}$.

$$D_{\mathbf{u},\mathbf{v}}(\pi) = \sum_{i}^{|\pi|} \phi(u_{\pi_1(i)}, v_{\pi_2(i)}) \tag{3}$$

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:

$$k_{GA}(\mathbf{u}, \mathbf{v}) = \sum_{\pi \in A(n,m)} e^{-D_{\mathbf{u},\mathbf{v}}(\pi)}, \tag{4}$$

where $A(n, m)$ denotes the set of all alignments between two time series of length n and m. Equation (4) can be rewritten by breaking up the alignment distances according to the local divergences: similarity function κ is induced by divergence ϕ:

$$k_{GA}(\mathbf{u}, \mathbf{v}) = \sum_{\pi \in A(n,m)} \prod_{i=i}^{|\pi|} e^{-\phi(u_{\pi_1(i)}, v_{\pi_2(i)})} \tag{5}$$

$$= \sum_{\pi \in A(n,m)} \prod_{i=i}^{|\pi|} \kappa\left(u_{\pi_1(i)}, v_{\pi_2(i)}\right), \tag{6}$$

where notation $\kappa = e^{-\phi}$ was introduced for the sake of simplicity. It has been argued that k_{GA} runs over the whole spectrum of the costs and gives rise to

a smoother measure than the minimum of the costs, i.e., the DTW (dynamic time warping) distance [10]. It has been shown in the same paper that k_{GA} is positive definite provided that $\kappa/(1+\kappa)$ is positive definite on $\mathcal{X}$. Furthermore, the computational effort is similar to that of the DTW distance; it is $\mathcal{O}(nm)$. Cuturi argued in [12] that global alignment kernel induced Gram matrix do not tend to be diagonally dominated as long as the sequences to be compared have similar lengths.

In our numerical simulations, we used local kernel $e^{-\phi_\sigma}$ suggested by Cuturi, where

$$\phi_\sigma(x,y) = \frac{1}{2\sigma^2}||x-y||^2 + \log\left(2 - e^{-\frac{||x-y||^2}{2\sigma^2}}\right). \tag{7}$$

2.3 Time-series Classification using SVM

Support Vector Machines (SVMs) are very powerful for binary and multi-class classification as well as for regression problems [7]. They are robust against outliers. For two-class separation, SVM estimates the optimal separating hyper-plane between the two classes by maximizing the margin between the hyper-plane and closest points of the classes. The closest points of the classes are called support vectors; the optimal separating hyper-plane lies at half distance between them.

In case of time-series classification, we are given sample and label pairs $\{(\mathbf{u}^{(i)}, l^{(i)})\}_{i=1}^{K}$ with $\left(\mathbf{u}^{(i)}, l^{(i)}\right) \in \left(\mathbb{R}^d\right)^{\mathbb{N}} \times \{-1, 1\}$. Here, for class '1' and for class '2' $l^{(i)} = 1$ and $l^{(i)} = -1$, respectively. We also have a feature map $\varphi : \left(\mathbb{R}^d\right)^{\mathbb{N}} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert-space. The kernel implicitly performs the dot product calculations between mapped points: $k(\mathbf{u}, \mathbf{v}) = \langle \varphi(\mathbf{u}), \varphi(\mathbf{v}) \rangle_{\mathcal{H}}$. The support vector classification seeks to minimize the cost function

$$\min_{w,b,\xi} \frac{1}{2}\|w\|_{\mathcal{H}}^2 + C\sum_{i=1}^{K} \xi_i \tag{8}$$

subject to the constraints

$$l^{(i)}\left(\left\langle w, \varphi\left(\mathbf{u}^{(i)}\right)\right\rangle_{\mathcal{H}} + b\right) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \tag{9}$$

where ξ_i-s are the so-called slack variables that generalize the original SVM concept with separating hyper-planes to soft-margin classifiers that have outliers that can not be separated.

We used multi-class classification, where decision surfaces are computed for all class pairs, i.e., for classes one has decision surfaces and then applies a voting strategy for decisions. We used the one-against-one procedure.

2.4 Multi-class Classification of Time Series Using Structured Sparsity

To tackle our multi-class AU learning problem, we exemplify structured-sparse coding defined for Euclidean spaces [23] to time series: (i) the occurrence is

captured by a non-overlapping group structure ($\mathcal{G}$), (ii) the underlying similarity of time series is handled by the global alignment kernel (k_{GA}, see Section 2.2).

Formally, let us assume that we are given a $k = k_{GA}$ kernel [34] on $\left(\mathbb{R}^d\right)^{\mathbb{N}}$, the set of d-dimensional time-series. Since k is a kernel there exists a feature mapping

$$\varphi : \left(\mathbb{R}^d\right)^{\mathbb{N}} \to \mathcal{H} \tag{10}$$

to a Hilbert space $\mathcal{H}$, where k represents an inner product

$$k(\mathbf{u}, \mathbf{v}) = \langle \varphi(\mathbf{u}), \varphi(\mathbf{v}) \rangle_{\mathcal{H}}, \quad \forall (\mathbf{u}, \mathbf{v}) \in \left(\mathbb{R}^d\right)^{\mathbb{N}} \times \left(\mathbb{R}^d\right)^{\mathbb{N}}.$$

Let us also assume that we have a $D = [\varphi(\mathbf{d}_1), \ldots, \varphi(\mathbf{d}_M)]$ dictionary and a $\mathcal{G}$ partition on $\{1, \ldots, M\}$, i.e., $\forall G_i, G_j \in \mathcal{G}$: $G_i \cap G_j = \emptyset$ $(i \neq j)$ and $\cup_{G \in \mathcal{G}} = \{1, \ldots, M\}$.

We aim to approximate an observation $\mathbf{x} \in \left(\mathbb{R}^d\right)^{\mathbb{N}}$ using the D dictionary taking into account the group-structure $\mathcal{G}$:

$$J(\boldsymbol{\alpha}) = \frac{1}{2} \left\| \varphi(\mathbf{x}) - \sum_{i=1}^{M} \varphi(\mathbf{d}_i)\alpha_i \right\|_{\mathcal{H}}^2 + \kappa \Omega(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha}}, \tag{11}$$

where

$$\Omega(\boldsymbol{\alpha}) = \sum_{G \in \mathcal{G}} \|\boldsymbol{\alpha}_G\|_q \quad (q \geq 1). \tag{12}$$

The occurrence of events is encoded by the Ω group-structure inducing regularizer: each $G \in \mathcal{G}$ corresponds to the activity of one type of event, by the application of Ω few events are favored. We used the so-called-ℓ_1/ℓ_2 norm ($q = 2$). Note that using $q = 1$ leads to ℓ_1-norm with no group sparsity effects. $\kappa > 0$ is a regularization parameter describing the trade-off between the two cost terms. In the sequel, optimization task (11) will be referred to as the kernel structured sparse coding problem of time series, or shortly KSS.

By applying the kernel trick, optimization of objective (11) is equivalent to

$$J(\boldsymbol{\alpha}) = \left(\frac{1}{2} \boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - \mathbf{k}^T \boldsymbol{\alpha} \right) + \kappa \Omega(\boldsymbol{\alpha}), \tag{13}$$

where $\mathbf{k} = [k(\mathbf{x}, \mathbf{d}_1); \ldots; k(\mathbf{x}, \mathbf{d}_M)] \in \mathbb{R}^M$ and $\mathbf{G} = [G_{ij}] = [k(\mathbf{d}_i, \mathbf{d}_j)] \in \mathbb{R}^{M \times M}$ is the Gram-matrix.

Equation (13) is a finite dimensional problem, which can be optimized, for example, by FISTA (fast iterative shrinkage-thresholding algorithm) [3]. Our experiments were based on the modification of the SLEP package [20]. The supplementary material provides additional details for the implementation (Online Resource 1).

Note: according to our numerical experiences on time series, it is often advantageous to apply normalization to the dictionary atoms and to the observations

$$\|\varphi(\mathbf{d}_i)\|_{\mathcal{H}} = 1, \quad (\forall i), \qquad\qquad \|\varphi(\mathbf{x})\|_{\mathcal{H}} = 1. \tag{14}$$

This can be carried out implicitly in the proposed approach by using the modified kernel

$$\bar{k}(\mathbf{u},\mathbf{v}) = \frac{k(\mathbf{u},\mathbf{v})}{\sqrt{k(\mathbf{u},\mathbf{u})}\sqrt{k(\mathbf{v},\mathbf{v})}} = \left\langle \frac{\varphi(\mathbf{u})}{\|\varphi(\mathbf{u})\|_{\mathcal{H}}}, \frac{\varphi(\mathbf{v})}{\|\varphi(\mathbf{v})\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}}. \tag{15}$$

Classification Using Structured Sparsity. In the multi-class case we investigated three different strategies for the classification of the input ($\mathbf{x}$) using the $\boldsymbol{\alpha}$ representation provided by the KSS method:

$$\hat{k} = \underset{k=1,\ldots,K}{\arg\max} \|\boldsymbol{\alpha}_{G_k}\|_2\,, \tag{16}$$

$$\hat{k} = \underset{k=1,\ldots,K}{\arg\min} \|\varphi(\mathbf{x}) - D_k\boldsymbol{\alpha}_{G_k}\|_{\mathcal{H}}\,, \tag{17}$$

$$\hat{k} = \underset{k=1,\ldots,K}{\arg\max} \left\| \varphi(\mathbf{x}) - \sum_{i=1;i\neq k}^{K} D_i\boldsymbol{\alpha}_{G_i} \right\|_{\mathcal{H}}, \tag{18}$$

where $D = [D_1,\ldots,D_K]$ is the dictionary partitioned according to $\mathcal{G}$ group structure ($K = |\mathcal{G}|$).

Intuitively, the first strategy [(16)] selects the group with the highest activity in the hidden representation $\boldsymbol{\alpha}$. The second one [(17)] chooses the group that minimizes the reconstruction error, the third one [(18)] selects the group, whose complement has the highest reconstruction error.

3 Experiments

3.1 Datasets

One motion gesture and two face datasets were used for evaluation. For gestures, we used the 6D Motion Gesture Database [9]. For emotion-specified expressions, we used the Cohn-Kanade Extended Facial Expression (CK+) Database [24]. For AU-labeled facial expressions, we used a subset of the more challenging Group Formation Task dataset [32]. See Table 1 for descriptive statistics of the datasets.

Table 1. Database statistics

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)

6D Motion Gesture Database. The 6D Motion Gesture Database (6DMG) [9] contains comprehensive motion data, including the 3D position, orientation, acceleration, and angular speed, for a set of different motion gestures performed by different users. The dataset composed of three subsets: motion gestures, air-handwriting and air-fingerwriting. In our experiments we used the air-handwriting set. WorldViz PPT-X4 was used as the optical tracking system, which tracks infrared dots. As for inertial sensors, the Wii Remote Plus embedded MEMS accelerometers and gyroscope were used. Overall, the tracking device provided both explicit and implicit 6D spatio-temporal information, including the position, orientation, acceleration, and angular speed. To eliminate allographs or different stroke orders, the subjects were instructed to follow certain "stroke order" for each character. The database contains 26 motion characters (uppercase A to Z) from 22 participants. Each character is repeated 10 times for every subject.

Cohn-Kanade Extended Dataset. The Cohn-Kanade Extended Facial Expression (CK+) Database [24] was developed for automated facial image analysis. CK+ together with an earlier version [18] is one of the most widely used testbeds for this purpose. The database contains 123 different subjects with 593 frontal image sequences. Each sequence begins with a neutral expression and ends at the apex of an emotion expression. A total of 327 image sequences have validated annotation for seven universal emotions (anger, contempt, disgust, fear, happy, sad and surprise). These image sequences were used in the current study. For each image sequence, 3D landmarks and shape parameters were obtained using the ZFace tracker.

Group Formation Task Corpus (spontaneous). For Action Unit onset classification task we used the more challenging Group Formation Task (GFT) corpus [32]. The corpus was built to evaluate the socioemotional effects of alcohol. It consist of 36 minutes of casual social interaction in 240 3-person groups of previously unacquainted young adults. Groups were randomly assigned to one of three conditions: alcoholic drinks, placebo beverages, and non-alcohol control beverages. Participants were recorded using three hardware synchronized cameras while seated around a circular table. Mean head pose is mostly frontal but moderate head pose variation and self-occlusion are common as subjects turn toward or away from each other. Facial AU occur during speech and in both additive and non-additive combinations. In the latter, individual AU modify each other's appearance. A sample video clip is available in the supplement (Online Resource 2).

Highly trained and certified FACS coders at the Affect Analysis Group fully FACS coded 3 minutes of video from each of 50 subjects [14]. This subset, denoted as GFT50, consists of 235,032 frames, annotated with 34 AUs. We selected 12 AUs for this evaluation. All AU had high inter-observer reliability as quantified by coefficient kappa (see Table 2), even the ones with high degree of skew, which can attenuate measures of agreement [16].

Table 2. FACS reliability on the GFT50 dataset. The skew column shows the imbalance ratio of the negative and positive ground truth labels.

AU	FACS Name	Cohen's κ	skew	AU	FACS Name	Cohen's κ	skew
1	Inner Brow Raiser	0.936	12.0	12	Lip Corner Puller	0.911	1.8
2	Outer Brow Raiser	0.857	7.4	14	Dimpler	0 895	0.7
4	Brow Lowerer	0.912	160.3	16	Lower Lip Depressor	0.858	52.5
7	Lid Tightener	0.942	2.6	17	Chin Raiser	0.833	2.4
10	Upper Lip Raiser	0.961	1.8	20	Lip Stretcher	0.914	91.4
11	Nasolabial Deepener	0.971	5.4	22	Lip Funneler	0.798	152.2

3.2 Time-Series Dictionary Building

For the experiments on the gesture dataset, we used the time-series data provided within the dataset. This data includes the position (3D), velocity (3D), orientation (4D), acceleration (3D), and angular speed (3D) of the movements. We evaluated the effectiveness of these modalities separately.

For the experiments using the facial expression datasets we formed time-series from shape- and appearance-based features. In the case of the CK+ dataset, we tracked the video sequences with the ZFace tracker and built the time-series from the PCA coefficients of the 3D PDM (parameter $\mathbf{q}$ in (2)). Illustratively, this is the compressed representation of the 3D landmark locations without rigid head movements. Our PDM contains 56 non-rigid parameters.

In the case of GFT50 dataset (action unit classification task), we used appearance features beside the shape. Using the 3D information first we acquired a canonical view (without rigid movement) of the tracked faces and then extracted SIFT descriptors around the markers. We used 49 landmarks (excluding the jawline points), thus we had a 6272 dimensional representation. We compressed the data to 166 dimension by means of PCA. We retained 90% of the variance. We formed time-series from this compressed holistic SIFT representation.

Since all three datasets come with meta-data (i.e., motion character, emotion or AU labels) we framed the classification task as a supervised learning problem and formed the non-overlapping groups structures in the dictionary according to the labels. In all experiments we employed a leave-one-subject out cross-validation: we removed all time-series instances of a given subject from the dictionary for testing and used the rest of the atoms for the training. For parameter selection we applied the same protocol on the training set in a nested scheme using the remaining subjects. We repeated this procedure for each subjects.

3.3 Gesture Classification on 6DMG

In this set of experiment we studied the structured sparsity method on the 6DMG motion database. 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.

Table 3. The character error rates (CER) of motion character recognition using single modalities. The different attributes in the columns: position (P), velocity (V), acceleration (A), angular velocity (W), orientation (O). The best result for each modality is denoted with bold letters. The best result for each method is denoted with underline.

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

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. The results and comparison with Hidden Markov Model (HMM) are summarized in Table 3. All methods achieved the best result using the 3D position data from all the available modalities. The KSS method outperformed both the HMM and the SVM techniques.

3.4 Emotional Expression Classification on CK+

In this set of experiments we studied the structured sparsity method on the CK+ dataset. We measured the performances of the methods for emotion recognition.

First, we tracked facial expressions with the ZFace tracker and annotated all image sequences starting from the neutral expression to the peak of the emotion. The tracker estimates the rigid and non-rigid transformations. We removed the rigid ones from the faces and represented the sequences as multi-dimensional time-series built from the 56 non-rigid shape parameters (parameter $\mathbf{q}$ in Eq.(2)).

We calculated Gram matrices using the GA kernels and performed leave-one-subject out cross validation to maximally utilize the available set of training data. We searched for the best parameter (σ of GA kernel) between 2^{-5} and 2^{10} on a logarithmic scale with equidistant steps and selected the parameter having the lowest mean classification error. The SVM regularization parameter (C) was searched within 2^{-5} and 2^{5} and the KSS regularization parameter (κ) was searched within 2^{-5} and 2^{1} in a similar fashion.

The result of the classification using the different voting strategies is shown in Table 4.a. Performance scores show that the time-series kernel SVM and the KSS method perform equally well on this task, achieving F_1 scores of 0.935 and 0.932, respectively.

For detailed comparisons with other sparse and non-sparse methods, see Table 5. We report both classification accuracy (Acc) and Area Under ROC Curve (AUC) values. The time-series kernel SVM outperforms all the non-sparse methods, including frame based and fixed length dynamic techniques. The KSS method outperforms frame based sparse methods that utilize shape or Gabor features. We note that in our experiments both the time-series kernel SVM and the KSS method rely on simple shape features. An interesting comparison can be made with Jeni et al. [17], where 3D CLM based shape features were used, but

Table 4. Classification results on (a) CK+ and (b) GFT50 datasets

(a)

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

(b)

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

Table 5. Comparisons with different sparse and non-sparse methods on CK+. We include (1) Frame level methods, (2) Fixed length spatio-temporal methods and (3) Varying length time-series methods.

	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**	-	-	-

the experiments were limited to individual frames without respect to temporal organization. By using the precisely aligned temporal information, the time-series kernel SVM wins by a considerable (11.1%) margin. Another important comparison concerns the study of Zafeiriou et al. [42], where sparsity was used on shape features, however only on frame level. Both our time-series kernel SVM and our KSS achieve more than 5% higher accuracy, indicating that temporal information is somewhat more important than the sparse representation in this case.

3.5 Action Unit Onset Classification on GFT50

Encouraged by the results of the previous experiment, we decided to test the methods for AU onset classification in order to estimate performance in the early phase of facial events.

We tracked facial expressions and extracted time series between 5 and 10 frames from AU onsets and trained kernel SMVs for one-vs-one AU classification and KSS with the three different voting strategies. Figure 2 shows the classification performance.

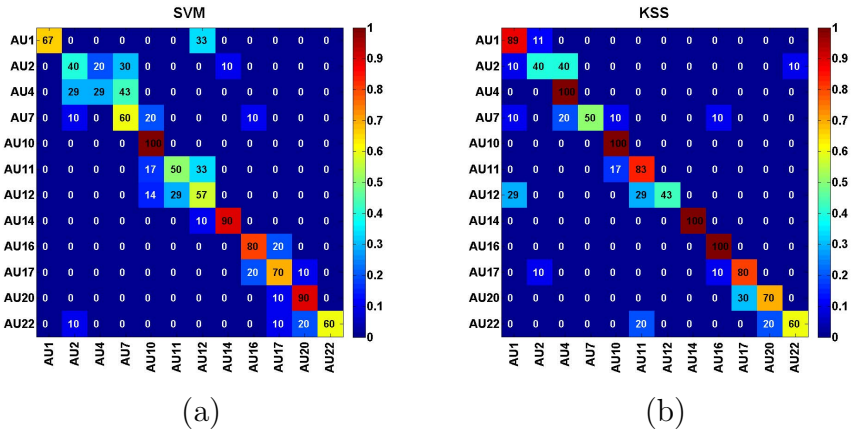

Fig. 2. Confusion matrices for early AU onset classification on the GFT50 corpus using (a) time-series kernel SVM and (b) Kernel Structured Sparsity methods

According the figures, from the three different voting strategies for the structured sparsity method the first [(16)] performed the best: its performance is superior to the multi-class time-series kernel SVM for the AU estimation task by a large (10%) margin. See Table 4.b for the comparison.

4 Conclusions

Facial expression estimation is a challenging problem for computer vision. Progress has been enabled by two factors. Large, annotated databases developed over the years were the means of developing precise texture and shape based descriptors and models. In the meantime, novel, efficient, and fast kernel-based similarity measures have been developed that can compare spatio-temporal patterns subject to time warping. We have shown that the efficiency of kernel methods can be further enhanced by sparse structured algorithms, at least for the gesture and facial expression datasets that we studied. These novel, structured methods approximate spatio-temporal patterns by a few groups of such patterns. The method achieves density within groups due to the squared norm and sparsity between them using the group-structure inducing regularizer.

Our present method has two specific features.

Contrast enhancement. Structured sparsification is used for contrast enhancement. The best group then is selected using a voting strategy.

Implicit reconstruction in feature space. By taking into account spatio-temporal similarities using time-series kernels, the reconstruction can be carried out implicitly; the representation of the input is mixed from the representations of dictionary elements belonging to a few groups.

We tested this method for detection of hand gestures, holistic emotions, and action units. For each, the method successfully represented concurrent processes

(see, (11) and the subsequent explanation) (e.g., different AUs). The efficiency and limitations of this method in other types of data is a research question.

Classical sparse models try to select a few elements of the representation such that the corresponding samples approximate the input. The error of the estimation is then computed and it can drive correcting steps. Error based correction makes it a feedback approximation. Reconstruction in feature space changes this algorithmic procedure. The input is compared with the samples, and the vector created from the individual similarity values is sparsified via the minimization of a quadratic expression, which results in a feedforward procedure. This procedure enables the straightforward application of sophisticated kernels and structured sparsification simultaneously.

For both AU and hand gesture classification, we found that structured sparse methods with reconstruction in feature space (KSS) out-performed multi-class SVM. This finding applied for all three variants of KSS, with few differences among variants. For holistic expressions, differences between KSS and multi-class SVM were negligible. The lack of effective differences for holistic expressions may be due in part to ceiling effects. Detection of holistic expressions by both KSS and multi-class SVM approached 100%. An additional factor may be that the number of classes for holistic expressions was relatively small. Because SVM scales quadratically as the number of classes increases, the relatively small number of holistic expression classes may have been insufficient to attenuate performance relative to KSS. This latter possibility is a research question. In summary, by combining temporal alignment and structured sparse reconstruction, KSS was comparable to multi-class SVM for holistic expressions and achieved marked advantage in event classification for both AU and hand gesture.

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 National Development Agency, Hungary (grant agreement Research and Technology Innovation Fund. EITKIC 12.); and the Gatsby Charitable Foundation.

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. Baraniuk, R.G., Cevher, V., Duarte, M.F., Hegde, C.: Model-based compressive sensing. IEEE Transactions on Information Theory 56, 1982–2001 (2010)
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2, 183–202 (2009)
4. Bousmalis, K., Morency, L.P., Pantic, M.: Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In: Automatic Face and Gesture Recognition, pp. 746–752 (2011)
5. Bousmalis, K., Zafeiriou, S., Morency, L.P., Pantic, M.: Infinite hidden conditional random fields for human behavior analysis. IEEE Transactions on Neural Networks and Learning Systems 24(1), 170–177 (2013)

6. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The convex geometry of linear inverse problems. Foundations of Computational Mathematics 12(6), 805–849 (2012)
7. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), http://www.csie.ntu.edu.tw/~cjlin/libsvm
8. Chen, M.: Universal Motion-Based Control and Motion Recognition. Ph.D. thesis, Georgia Institute of Technology (2013)
9. Chen, M., AlRegib, G., Juang, B.H.: 6dmg: A new 6d motion gesture database. In: 3rd Multimedia Systems Conference, MMSys 2012, pp. 83–88. ACM, New York (2012)
10. Cuturi, M., Vert, J.P., Birkenes, Ø., Matsui, T.: A kernel for time series based on global alignments. In: International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 413–416 (2007)
11. Cuturi, M.: Fast global alignment kernels. In: International Conference on Machine Learning (ICML), pp. 929–936 (2011)
12. Cuturi, M.: Fast global alignment kernels. In: International Conference on Machine Learning, pp. 929–936 (2011)
13. Zhou, F., de la Torre, F., Hodgins, J.K.: Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(3), 582–596 (2013)
14. Girard, J., Cohn, J.: Ground truth FACS action unit coding on the group formation task. Tech. rep., University of Pittsburgh (2013)
15. Huang, J., Zhang, T., Metaxas, D.: Learning with structured sparsity. Journal of Machine Learning Research 12, 3371–3412 (2011)
16. Jeni, L., Cohn, J., de la Torre, F.: Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 245–251 (September 2013)
17. Jeni, L.A., Lőrincz, A., Nagy, T., Palotai, Z., Sebők, J., Szabó, Z., Takács, D.: 3d shape estimation in video sequences provides high precision evaluation of facial expressions. Image and Vision Computing 30(10), 785–795 (2012)
18. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Automatic Face and Gesture Recognition, pp. 46–53 (2000)
19. Koltchinskii, V., Yuan, M.: Sparse recovery in large ensembles of kernel machines on-line learning and bandits. In: COLT, pp. 229–238 (2008)
20. Liu, J., Ji, S., Ye, J.: SLEP: Sparse learning with efficient projections (2010), http://www.public.asu.edu/~jye02/Software/SLEP/
21. Long, F., Wu, T., Movellan, J.R., Bartlett, M.S., Littlewort, G.: Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. Neurocomputing 93, 126–132 (2012)
22. Lőrincz, A., Jeni, L.A., Szabó, Z., Cohn, J.F., Kanade, T.: Emotional expression classification using time-series kernels. In: Computer Vision and Pattern Recognition Workshops (CVPRW), Portland, OR (2013)
23. Lu, Y.M., Do, M.N.: A theory for sampling signals from union of subspaces. IEEE Transactions on Signal Processing 56(6), 2334–2345 (2008)
24. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)

25. Mahoor, M., Zhou, M., Veon, K.L., Mavadati, S., Cohn, J.: Facial action unit recognition with sparse representation. In: Automatic Face Gesture Recognition and Workshops, pp. 336–342 (March 2011)
26. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision 60(2), 135–164 (2004)
27. Obozinski, G., Wainwright, J., Jordan, M., Support, M.I.: union recovery in high-dimensional multivariate regression. Annals of Statistics 39(1), 1–17 (2011)
28. Ekman, P., Friesen, W., Hager, J.: Facial action coding system: Research nexus. Network Research Information, Salt Lake City (2002)
29. Ekman, P., Friesen, W.F.: Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
30. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 26(1), 43–49 (1978)
31. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision 91(2), 200–215 (2011)
32. Sayette, M., Creswell, K., Dimoff, J., Fairbairn, C., Cohn, J., Heckman, B., Kirchner, T., Levine, J., Moreland, R.: Alcohol and group formation: a multimodal investigation of the effects of alcohol on emotion and social bonding. Psychological Science 23(8), 869–878 (2012)
33. Sikka, K., Wu, T., Susskind, J., Bartlett, M.: Exploring bag of words architectures in the facial expression domain. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part II. LNCS, vol. 7584, pp. 250–259. Springer, Heidelberg (2012)
34. Steinwart, I., Christmann, A.: Support Vector Machines. Springer (2008)
35. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288 (1996)
36. Tropp, J.A., Wright, S.J.: Computational methods for sparse solution of linear inverse problems. In: Proceedings of the IEEE Special Issue on Applications of Sparse Representation and Compressive Sensing, pp. 948–958 (2010)
37. Valstar, M.F., Pantic, M.: Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In: Lew, M., Sebe, N., Huang, T.S., Bakker, E.M. (eds.) HCI 2007. LNCS, vol. 4796, pp. 118–127. Springer, Heidelberg (2007)
38. Wu, T., Bartlett, M., Movellan, J.R.: Facial expression recognition using Gabor motion energy filters. In: Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 42–47 (2010)
39. Xiong, X., de la Torre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (June 2013)
40. Yang, P., Liu, Q., Metaxas, D.N.: Boosting encoded dynamic features for facial expression recognition. Pattern Recognition Letters 30(2), 132–139 (2009)
41. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68(1), 49–67 (2006)
42. Zafeiriou, S., Petrou, M.: Sparse representations for facial expressions recognition via l1 optimization. In: Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 32–39 (June 2010)