Characteristic Kernels on Structured Domains Excel in Robotics and Human Action Recognition

Somayeh Danafar [1], Arthur Gretton [2], and Jürgen Schmidhuber [1]

[1]Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
Galleria 2, 6928 Manno-Lugano, Switzerland
[2] Carnegie Mellon University, Pittsburgh, USA,
and MPI for Biological Cybernetics, Tübingen, Germany
{somayeh, juergen}@idsia.ch, arthur.gretton@gmail.com

Abstract. Embedding probability distributions into a sufficiently rich (characteristic) reproducing kernel Hilbert space enables us to take higher order statistics into account. Characterization also retains effective statistical relation between inputs and outputs in regression and classification. Recent works established conditions for characteristic kernels on groups and semigroups. Here we study characteristic kernels on periodic domains, rotation matrices, and histograms. Such structured domains are relevant for homogeneity testing, forward kinematics, forward dynamics, inverse dynamics, etc. Our kernel-based methods with tailored characteristic kernels outperform previous methods on robotics problems and also on a widely used benchmark for recognition of human actions in videos.

Keywords: Characteristic kernels, Locally compact Abelian groups, Rotation matrices, Semigroups, Recognition of human actions in videos

1 Introduction

Kernel methods solve difficult non-parametric problems by embedding data points in higher-dimensional reproducing kernel Hilbert spaces (RKHS). This property makes kernel methods useful and strong tools to be used in different tasks. They were successfully applied to a wide range of learning tasks such as regression and classification [16]. Recent studies focused on mapping random variables into RKHS to collect linear statistics in RKHS which in turn were used to derive their meaning in the original space [8], [19], [20]. When the embedding is injective, the RKHS is said to be *characteristic* [5]. Such mappings allow for testing whether two distributions coincide [8],[9], or for finding the most predictive subspace in regression [6]. The most predictive (effective) subspace in regression is obtained by isolating the features that capture the statistical relationship between inputs and targets.

Characteristic kernels are defined on non-compact and complex domains. Sriperumbudur et al. [20] showed that a continuous shift-invariant $\mathbb{R}$-valued positive definite kernel on $\mathbb{R}^n$ is characteristic if and only if the support of its Fourier transform is the entire $\mathbb{R}^n$. Fukumizu et al. [7] extended Fourier analysis to groups and semigroups and obtained necessary conditions for defining characteristic kernels on spaces other than $\mathbb{R}^n$.

The main contribution of this paper is empirical evaluation of characteristic kernels. We investigate characteristic kernels on structured domains (groups/ semigroups) for various kernel-based methods: Maximum Mean Discrepancy (MMD) [8], [9] as a non-parametric hypothesis test, Support Vector Regression with ϵ-insensitive loss function (ϵ-SVR) [18], Gaussian Process Regression (GPR) [14] as a non-parametric regression method, and Support Vector Machines (SVM) [16] to classify human actions in videos. We provide experimental evidence that these kernel-based methods with appropriate kernels lead to significant performance gains.

Section 2 briefly reviews kernel-based methods. Section 3 introduces novel characteristic kernels on periodic data, the orthogonal group SO(3), and histograms. Section 4 experimentally confirms their theoretical advantages: we obtain state-of-the-art results in homogeneity testing, forward kinematics, forward dynamics, inverse dynamics, and recognition of human actions in videos.

2 Kernel-Based Learning Methods

In this section we briefly review some of kernel-based methods that we use to investigate our characteristic kernels.

2.1 A Non-Parametric Statistical Test

One basic statistic on Euclidean space is the *mean*. By embedding the distributions into RKHS, the corresponding factor is the *mean element*. The distance between mapped mean elements is known as Maximum Mean Discrepancy (MMD) [8], [9]. The definition of MMD is given in the following theorem:

Theorem 1. *Let $(\mathcal{X}, \mathcal{B})$ be a metric space, and let P, Q be two Borel probability measures defined on $\mathcal{X}$. Then $P = Q$ if and only if* $\mathrm{MMD}(P, Q) = 0$, where

$$\begin{aligned}\mathrm{MMD}(P,Q) &:= \parallel \mu_P - \mu_Q \parallel_{\mathcal{H}} \\ &= \parallel E_P[k(x,.)] - E_Q[k(y,.)] \parallel_{\mathcal{H}} \\ &= (E_{x,x' \sim P}[k(x,x')] + E_{y,y' \sim Q}[k(y,y')] - 2E_{x \sim P, y \sim Q}[k(x,y)])^{\frac{1}{2}} \qquad (1)\end{aligned}$$

One application of MMD is homogeneity testing, which tests whether the samples were drawn from different distributions. We compare MMD to another two-sample test suited for periodic distributions, namely, the Uniform Scores Test (UST) [4]. UST is not a kernel-based method.

Uniform Scores Test (UST) UST [4] is a two-sample test which tests whether distributions of circular data coincide. In UST each distribution is represented by a radius. The null hypothesis is rejected if the summation of radii is too large. Here we define UST more precisely.

Suppose we have n_i samples where $i = 1, 2, .., r$. We treat sample $n_1 = \{\theta_1, ..., \theta_n\}$ as linear data, re-arrange them in ascending order, and assign rank r_i to each θ_i. The *circular rank* of θ_i is then defined as $\gamma_i = 2\pi r_i/n$, for $i = 1, ..., n$. We denote γ_i as the *uniform score* corresponding to θ_i. We take all $N = n_1 + ... + n_r$ data values as a single sample and calculate their circular ranks. Let γ_{ij} denote the circular rank of θ_{ij} among all the data. For each sample $n_i, i = 1, ..., r$, we calculate

$$C_i = \sum_{j=1}^{n_i} cos\gamma_{ij}, S_i = \sum_{j=1}^{n_i} sin\gamma_{ij} \tag{2}$$

and hence the test statistics

$$W_r = 2\sum_{i=1}^{r}(C_i^2 + S_i^2)/n_i \tag{3}$$

If W_r is too large, we reject the null hypothesis that the distributions are identical.

2.2 Non-parametric Regression Methods for Model Learning

The task of regression is to learn the input/target mapping, to predict target values for query inputs.

Support Vector Regression with ϵ-Insensitive Loss function (ϵ-SVR) The goal of ϵ-SVR regression is to find a mapping function $f(x)$ which for each training input x deviates from its target by at most ϵ, and simultaneously is as flat as possible. According to [19], $f(x)$ is

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i, x) + b. \tag{4}$$

where $K(x_i, x) = \phi(x_i)^T\phi(x)$, and i ranges over the training points. The solution of a quadratic optimization problem determines the quantities α_i^*, α_i, and b.

Gaussian Processes Regression (GPR) Gaussian Process Regression (GPR) [14] uses a linear model to find a latent function $f(x)$. Uncertainty is modeled probabilistically by:

$$f \sim N(0, \varPhi\varSigma\varPhi^T) \sim N(0, K) \tag{5}$$

where matrix $\varPhi$ describes transformation columns $\phi(x)$ for all cases in the training set, $\varSigma$ is the covariance matrix of the weights, and K is a positive semidefinite matrix with elements $K_{i,j} = k(x_i, x_j)$ for some covariance function $k(., .)$.

2.3 Classification: Support Vector Machines

Consider the problem of separating the training set into two classes. If we assume that the two classes can be separated by a hyperplane $w.x + b = 0$ in some space $\mathcal{H}$, and that we have no prior knowledge about the data distribution, then the optimal hyperplane maximizes the margin [16]. Optimal values for w and b can be found by solving a constrained minimization problem, using Lagrange multipliers $\alpha_i (i = 1, .., l)$. The classifier is defined as:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right) \tag{6}$$

where K is the kernel mapping data points to RKHS $\mathcal{H}$, and α_i and b are found using an SVC learning algorithm. Those x_i with nonzero α_i are called the *support vectors.*

3 Characteristic Kernels on Structured Domains

Characteristic kernels were defined on non-compact domain like entire $\mathbb{R}^n$. Sriperumbudur et al. [20] showed that if and only if the support of Fourier transform of a shift invariant positive definite kernel is the entire $\mathbb{R}^n$, this kernel is characteristic. A question that naturally arises is whether characteristic kernels can be defined on spaces besides $\mathbb{R}^n$. Several such domains constitute topological groups/semigroups. Fukumizu et al. [7] based on extensions of Fourier analysis to groups and semigroups established necessary and sufficient conditions of introducing characteristic kernels. Our main contribution in this paper is to study these characteristic kernels defined by their algebraic structure and assess them in relevant applications. For the sake of this purpose, thanks to the established conditions and theorems by Fukumizu et al. [7] we define our proper characteristic kernels. We investigate characteristic kernels on Locally Compact Abelian (LCA) groups (periodic data), Compact Groups (rotation matrices), and Abelian Semigroups (histogram-based data). In this section we clarify our characteristic kernels, thereafter relevant experiments and evaluations will be discussed in section 4.

3.1 Shift Invariant Characteristic Kernels on LCA groups

Periodic domains are examples of Locally Compact Abelian groups which we consider in this study. To define our proper characteristic kernels on periodic domains, we use Theorems 7 and 8 of [7] which describe necessary and sufficient conditions for kernels on LCA groups to be characteristic, as well as Corollary 9 of [7] on the multiplication of shift-invariant characteristic kernels, which is again a characteristic kernel. Our novel characteristic kernels on periodic domains are:

1. $k_1(x, y) = \prod_{i=1}^{l} (\pi - (x_i - y_i)_{mod\ 2\pi})^2$,
2. $k_2(x, y) = \prod_{i=1}^{l} (\cosh(\pi - (x_i - y_i)_{mod 2\pi})$,

3. $k_3(x, y) \;=\; \prod_{i=1}^{l}(-\log(1 - 2\alpha\cos(x_i - y_i) + \alpha^2),$
4. $k_4(x, y) \;=\; \prod_{i=1}^{l}(1 - \alpha^2)/(1 - 2\alpha\cos(x_i - y_i) + \alpha^2),$

where l denotes the input dimension. Periodic domains are relevant for two-sample testing, and in regression tasks like forward kinematics, forward dynamics, and inverse dynamics. In the case of forward dynamics besides periodic data we have torques which do not belong to periodic domain. We work with the following justified characteristic kernels in that case:

1. $k_5(x, y) \;=\; \prod_{i=1}^{m}(\pi \;-\; (x_i - y_i)_{mod\ 2\pi})^2 \cdot \text{Gaussian}(x_{m,..,l}, y_{m,..,l}),$
2. $k_6(x, y) \;=\; \prod_{i=1}^{m}(\cosh(\pi - (x_i - y_i)_{mod 2\pi}) \cdot \text{Gaussian}(x_{m,..,l}, y_{m,..,l}),$
3. $k_7(x, y) \;=\; \prod_{i=1}^{m}(-\log(1 - 2\alpha\cos(x_i - y_i) + \alpha^2) \cdot \text{Gaussian}(x_{m,..,l}, y_{m,..,l}),$
4. $k_8(x, y) \;=\; \prod_{i=1}^{m}(1-\alpha^2)/(1-2\alpha\cos(x_i-y_i)+\alpha^2)\cdot\text{Gaussian}(x_{m,..,l}, y_{m,..,l}).$

3.2 Characteristic Kernels on Compact Groups

Famous examples of non-Abelian topological groups are the ones consisting of matrices, such as the orthogonal group SO(3). According to Theorems 11 and 12 of [7], we define proper kernels on rotation matrices $\{A, B\} \in \mathbb{R}^3$. Let $\cos\theta = \frac{1}{2}Tr[B^{-1}A]$, and $0 \le \theta \le \pi$, we formulate the characteristic kernels as follows:

$$k_1(A, B) = \frac{1}{\sin\theta}\sum_{n=0}^{\infty}\frac{\sin((2n+1)\theta)}{(2n+1)^3} = \frac{\pi\theta(\pi-\theta)}{8\sin\theta}. \tag{7}$$

$$k_2(A, B) = \sum_{n=0}^{\infty}\frac{\alpha^{2n+1}\sin((2n+1)\theta)}{(2n+1)\sin\theta} = \frac{1}{2\sin\theta}\arctan\left(\frac{2\alpha\sin\theta}{1-\alpha^2}\right). \tag{8}$$

3.3 Characteristic Kernels on Abelian Semigroups

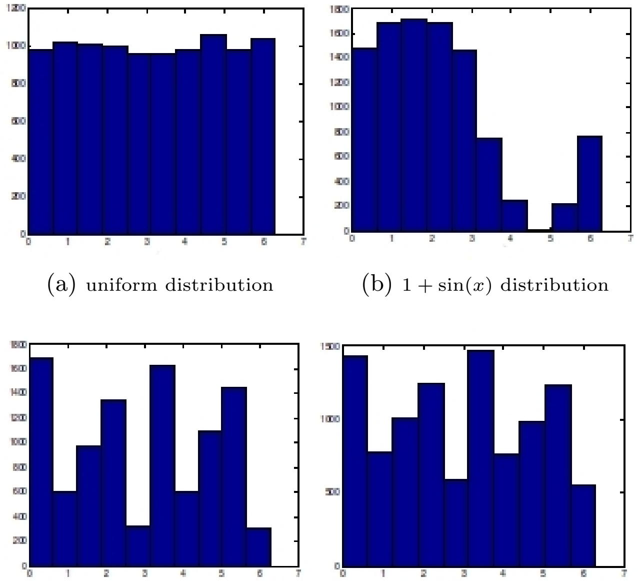
Now consider histograms as an example of Abelian semigroups such as $(\mathbb{R}_+^n, +)$. Theorems 13 and 14 of [7] obtain necessary and sufficient conditions for tailored kernels for histogram-based information. Let $a = (a_i)_{i=1}^n$ and $b = (b_i)_{i=1}^n$, $(a_i \ge 0, b_i \ge 0)$ be non-negative measures on n points. We use the following characteristic kernel:

$$k(a, b) = e^{-\beta\sum_{i=1}^{n}\sqrt{a_i+b_i}}. \tag{9}$$

where $\beta \ge 0$ and $\mathcal{X} \in \mathbb{R}$. Another tailored kernel for histogram-based data which is not a characteristic kernel is *Generalized Histogram Intersection (GHI)* kernel. In [1] GHI was introduced as a positive-definite kernel:

$$K_{\text{GHI}}(a, b) = \sum_{i=1}^{m}\min\{\mid a_i^{\beta} \mid, \mid b_i^{\beta} \mid\}, \quad (a, b) \in \mathcal{X}\times\mathcal{X} \tag{10}$$

We compare the results of these two kernels in human action classification task in section 4.4.



(a) uniform distribution (b) $1 + \sin(x)$ distribution

(c) $1 + \sin(3x)$ (d) $1 + \sin(6x)$

Fig. 1: (a)represents an example of circular data $[0, 2\pi)$ with uniform distribution, (b), (c), and (d) are periodic data with distribution $1 + \sin(\omega x)$ and ω equal to 1, 3, and 6 respectively. Higher perturbation frequencies make the perturbed distribution much closer to the uniform distribution, and the discrimination more difficult.

4 Experiments and Evaluations

Now we confirm theoretical advantages of characteristic kernels on various practical applications.

4.1 MMD for Two-Sample Testing

One application of MMD is for two-sample tests, which involve testing the null hypothesis $H_0 : P = Q$ versus $H_1 : P \neq Q$. Two-sample tests require a measure of distance between probabilities and a notion of whether this distance is statistically significant. Our MMD test determines the test threshold by the bootstrap procedure [8]. In this study we consider this application of MMD to compare two artificially generated distributions of periodic nature. Suppose we obtain the first sample from a uniform distribution P. The other sample is drawn from a perturbed uniform distribution $Q : 1+\sin(\omega x)$. For higher perturbation frequencies ω (where $1/\omega$ is smaller), it becomes harder to discriminate Q from P— see Figure 1.

Figure 2 shows the acceptance percentage of null hypothesis with MMD during 1000 runs with a user-defined significance level 0.05. The quality of MMD as a statistic depends on the richness of RKHS $\mathcal{H}$ which is defined by a measurable kernel. Characteristic kernels [5], [6] yield an RKHS for which probabilities have unique images. Here we use characteristic kernels $k_1,\ k_2,\ k_3$, and k_4 in MMD with $l = 1$ and hyper-parameter $\alpha = 0.9$ for kernels k_3 and k_4. MMD discriminated subtle changes between the distributions with the justified characteristic kernels on periodic domain. This can be seen from different acceptance percentage of H_0 in Figure 2. MMD has the best performance with k_4 which needs tuning a hyper-parameter α. We compared the result of MMD with UST. We observe that UST can not deal with subtle nonlinear changes in distributions. It gives true results when P and Q are either completely similar or dissimilar.

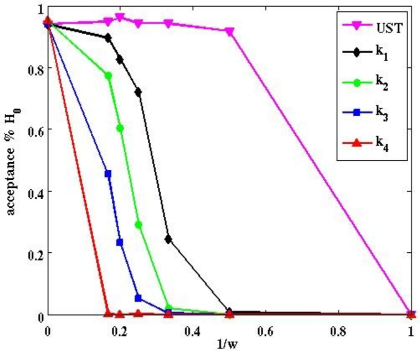


Fig. 2: Acceptance percentage of $H_0 : P = Q$ for MMD and UST, with user-defined significance level of 0.05 during 1000 runs, and $\frac{1}{\omega} = 0,\ \frac{1}{6},\ \frac{1}{5},\ \frac{1}{4},\ \frac{1}{3},\ \frac{1}{2}$, and 1. P is a uniform distribution of circular data, and Q is $1 + \sin(\omega x)$.

4.2 Applications of Regression

Given a training set of data points $\mathcal{D} = \{(x_1, y_1), ...(x_l, y_l)\}$ where the $x_i \in \mathbb{R}^n$ are inputs and the $y_i \in \mathbb{R}^1$ are the corresponding targets, the task of regression is to learn the input/target mapping, to predict target values for query inputs. Fukimuzu et al. [5], [6] showed that characterization allows us to derive a contrast function for estimation of the effective subspace. The effective subspace can help to retain the statistical relationship between x and y by isolating the features that capture this relation. We evaluated characteristic kernels $k_1,\ k_2,\ k_3$, and k_4 in forward kinematics and inverse dynamics for datasets with periodic nature. For forward dynamics problem, characteristic kernels $k_5,\ k_6,\ k_7$, and k_8 are used.

Forward kinematics Kinematics studies motion ignoring the forces which cause it. The forward kinematics of a revolute robot arm are described by the function $T = f(\theta, \phi)$, where θ is the vector of joint angles, ϕ are the parameters describing the kinematics of the arm, and T is the 4×4 homogeneous transformation matrix [2].

We use the 8 input *Kin* dataset (`http://www.cs.utoronto.ca/~delve/data/kin/desc.html`). It is generated from a realistic simulation of the forward kinematics of an 8 link all-revolute robot arm. The task is to predict the distance of the end-effector from a target, given the angular position of the 8 joints, the link twist angles, link lengths and link offset distances. Combinations of the following attributes are considered in datasets:

1. output : highly nonlinear (n) vs. fairly linear (f)

2. predicted value : medium noise (m) vs. high noise(h)

We use a training set of size 1024, 4096 test instances and the validation set of size 3072. The hyper-parameters α and σ in kernels k_3, k_4, and Guassian kernel respectively were tuned during the 5-fold leave-one-out cross validation procedure. Support vector regression with ϵ insensitive loss function ($\epsilon - SVR$) is used as our non-parametric regression method. A run consisted of model selection, training and testing, and the confidence interval over Mean Squared Errors (MSE) results are obtained over 10 runs. In this task the input dimension is 8. l is set to 8 in the formula of our characteristic kernels of section 3.1. In Figure 3, the results of $\epsilon - SVR$ with $\epsilon = 0.01$ on four datasets of 8-input Kin (Kin-8fm, Kin-8fh, Kin-8nm, and Kin-8nh) are depicted. Figure 3 demonstrates that tailored characteristic kernels on the LCA group work better than Gaussian kernel which is just characteristic. We compared our best results on the above datasets to the results given by GPR [14], K-Nearest Neighbor (K-NN), Linear Regression (LR), Multi-Layer Perceptrons (MLP) with single hidden layer and early stopping [14], and mixtures of experts trained by Bayesian methods (HME)[22]. The results reported in Table 1. Results of 22 methods (by Ghahramani) on the same datasets are available at `http://www.cs.toronto.edu/~delve/data/kin/desc.html`. The reported results show that GPR obtained better results than LR, as it captures the nonlinear relationship between data points by a Gaussian kernel and the affect of noise with probabilistic nature of the method. This draws the attention to our datasets which are generated by fairly linear and nonlinear movements of robot arm in combination with noise. Moreover the results of GPR in comparison with HME as another Bayesian based method is better which shows the superiority of kernel-based methods. The nonlinearities captured by MLP and GPR produced comparable results with better performance for GPR. Our results showed that $\epsilon - SVR$ and a tailored characteristic kernel on periodic data outperforms the other methods. This highlights the fact that in kernel-based methods selection of an appropriate kernel according to the nature of available data leads to significant performance gains. Our results with tailored characteristic kernels for periodic data confirm this fact.

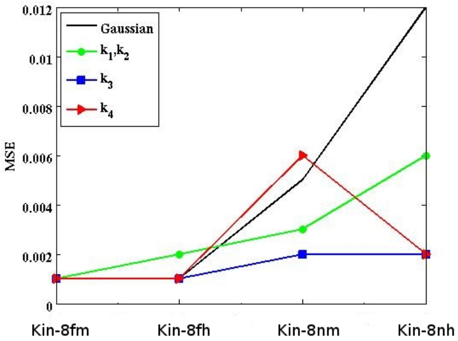


Fig. 3: The result of ϵ-SVR for forward kinematics task on Kin-8fm, Kin-8fh, Kin-8nm, and Kin-8nh datasets. The results of ϵ-SVR is based on characteristic kernels k_1, k_2, k_3, k_4 (section 3.1), and Gaussian kernel with $\epsilon = 0.01$, $\alpha = 0.9$, for k_3, and k_4, and $\sigma = 10$ for Gaussian kernel.

Table 1: Best results of ϵ-SVR with characteristic kernel k_4 in comparison with reported results on the Kin family of datasets. Our results are obtained with $\epsilon = 0.01$ in ϵ-SVR and the kernel k_4 with hyper-parameter $\alpha = 0.9$. Rounded standard deviation of MSEs are also reported.

Method	Kin-8fm	Kin-8fh	Kin-8nm	Kin-8nh
ϵ-SVR	0.001 ± 0.0001	0.001± 0.0001	0.002± 0.0001	0.002± 0.0001
GPR	0.25 ± 0.0001	0.02 ± 0.01	0.43 ± 0.1	0.1 ± 0.2
HME	0.26 ± 0.0001	0.03 ± 0.01	0.48 ± 0.3	0.28± 0.2
KNN	0.29 ± 0.0001	0.08 ± 0.01	0.65 ± 0.1	0.45 ± 0.2
LR	0.28 ± 0.0001	0.06 ± 0.01	0.65 ± 0.1	0.45 ± 0.2
MLP	0.26 ± 0.0001	0.03 ± 0.02	0.42 ± 0.01	0.1 ± 0.2

Forward dynamics To simulate robot control systems, forward dynamics computes joint accelerations and actuator torques, given position and velocity state [2]. We used the 8 input *Pumadyn* dataset at `http://www.cs.utoronto.ca/~delve/data/pumadyn/desc.html`. It was synthetically generated from a realistic simulation of the dynamics of a Puma560 robot arm. The task is to predict the angular acceleration of the robot arm links, given angular positions, velocities, torques. The combination of fairly linear and nonlinear movements of robot arm with unpredictability captured by medium or high amount of noise generate 4 datasets (Pdyn-8fm, Pdyn-8fh, Pdyn-8nm, and Pdyn-8nh). We used characteristic kernels k_5, k_6, k_7, and k_8 in this task. All the settings are like in the forward kinematics case. Figure 4 shows the justified characteristic kernels have better performance than Gaussian kernel. We compared our best results to those obtained by GPR [14], K-Nearest Neighbor (K-NN), Linear Regression

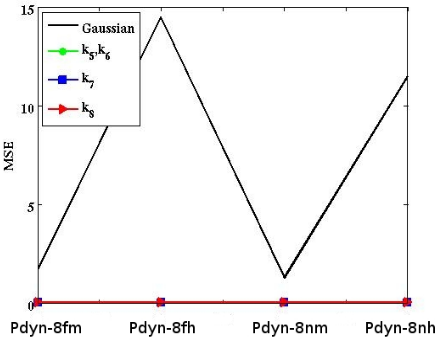


Fig. 4: The result of ϵ-SVR for forward dynamics task on Pdyn-8fm, Pdyn-8fh, Pdyn-8nm, and Pdyn-8nh. The results of ϵ-SVR is based on characteristic kernels k_5, k_6, k_7, and k_8 (section 3.1) with $\epsilon = 0.01$, $\alpha = 0.9$ for k_7, and k_8 respectively, and $\sigma = 10$ for Gaussian kernel.

Table 2: Best results of our ϵ-SVR with characteristic kernel k_8, as well as earlier reported results on the Pumadyn family of datasets. The results are obtained with $\epsilon = 0.01$ in ϵ-SVR and the kernel k_8 with hyper-parameter $\alpha = 0.9$. Rounded standard deviation of MSEs are also reported.

Method	P-8fm	P-8fh	P-8nm	P-8nh
ϵ-SVR	0.01±0.0001	0.01 ±0.0001	0.01 ±0.0001	0.01 ±0.0001
GPR	0.39 ± 0.001	0.05 ± 0.1	0.32± 0.01	0.03 ± 0.2
HME	0.41 ± 0.001	0.06 ± 0.1	0.37 ± 0.5	0.04 ± 0.3
KNN	0.41 ± 0.001	0.15 ± 0.1	0.52 ± 0.01	0.3 ± 0.1
LR	0.48 ± 0.001	0.08 ± 0.1	0.55 ± 0.01	0.48 ± 0.1
MLP	0.4 ± 0.001	0.06 ± 0.2	0.35 ± 0.01	0.033± 0.1

(LR), MLP with early stopping and single hidden layer [14], mixtures of experts trained by Bayesian methods (HME) [22] in Table 2. Results of 25 methods (by Ghahramani) are available at `http://www.cs.toronto.edu/~delve/data/pumadyn/desc.html`. Like the reported results in the forward kinematics case, the results of kernel based method GPR are better than those of linear Regression, and is better than HME method which is a Bayesian method. The results of GPR and MLP are comparable although the performance of GPR is better. The best outcome is for our ϵ-SVR method with justified characteristic kernels on datasets. ϵ-SVR captures the nonlinearity, and the relation of observations with tailored characteristic kernels.

Inverse Dynamics Finding sufficiently accurate dynamic models of rigid body equations in automatic robot control is difficult due to unmodeled nonlinearities, complex friction and actuator dynamics. Imprecise prediction of joint torques leads to poor control performance and may even damage the system. Learning more precise inverse dynamics models from measured data by regression is an interesting alternative. Here we compare $\epsilon - SVR$ and GPR as regression methods for computing inverse dynamics, which could be used for automatic robot control (e.g., [13]).

The inverse dynamic model [2] is given in the rigid-body formulation $u = M(q)\ddot{q} + F(q, \dot{q})$, where $q, \dot{q}, \ddot{q}$ are joint angles, angular velocities and angular accelerations of the robot. $M(q)$ denotes the inertia matrix and $F(q, \dot{q})$ the internal forces. Let us define the inverse dynamic model by $u = f(q, \dot{q}, \ddot{q})$; the regression task is to learn f.

We use the 7-DOF *SARCOS* anthropomorphic robot arm data `http://www.gaussianprocess.org/gpml/data`. Each observation in the data set consists of 21 input features (7 joint positions, 7 joint velocities, and 7 joint accelerations) and the corresponding 7 joint torques for the 7-DOF. There are two disjoint sets, one for training and one for testing. We use only 1100 examples of the training set for training, but the entire test set for testing. Results are shown in terms of normalized Mean Squared Errors (nMSEs) defined as MSE divided by target variance. Results of $\epsilon - SVR$ and GPR are shown in Figure 5. $\epsilon - SVR$ and GPR with tailored characteristic kernels work better than with the Gaussian kernel. Their results are comparable with slightly better performance in ϵ-SVR. The larger errors for the 5^{th} and 6^{th} DOF show that nonlinearities (e.g. hydrolic cables, complex friction) can not be approximated well using just the rigid body functions. This is an example of the difficulty of using an analytical model for control in practice.

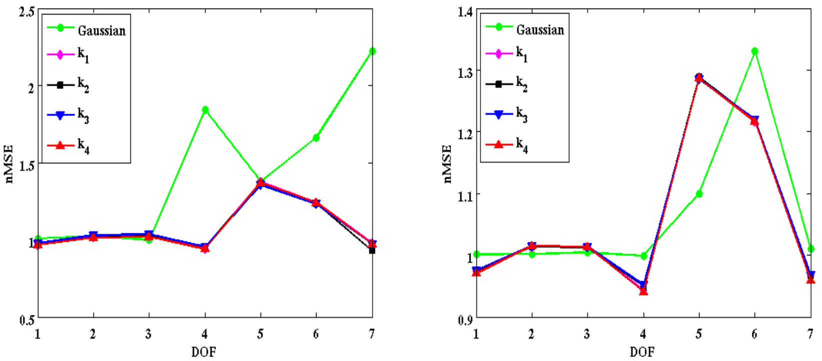
Yeung et al. [21] investigated different training sample sizes for GPR. They achieved the same result as reported in the current paper with GPR and Gaussian kernel over training set of size 1100 with the mean of nMSE = 1.06 and the standard deviation of nMSE = 0.12. They further improved their results by multi-task learning and reported the mean of nMSE =0.35 and the standard deviation of nMSE= 0.49 for multi-task GPR. From our improvement for both ϵ-SVR and GPR with tailored characteristic kernels in comparison with Gaussian kernel (Figure 5) we expect to see a performance boost in multi-task learning, but this is a topic of future work.

4.3 Rotation Matrices in Forward Kinematics

As mentioned before the task in forward kinematics is to find $T = f(\theta, \phi)$, where T a the 4×4 homogeneous rotation matrix (an example of SO(3) group). We considered the solution of the regression task with ϵ-SVR and the tailored characteristic kernels on rotation matrices of formula 7, 8, and Gaussian kernel on Kin-8nh dataset. We obtained the following results:

1. $k_1(A, B) \Rightarrow$ MSE = 0.009

2. $k_2(A, B)$ with $\alpha = 0.9 \Rightarrow$ MSE = 0.006



(a) Results of $\epsilon - SVR$ with tailored characteristic kernels on periodic domains and Gaussian kernel with $\epsilon = 0.01$ and $\alpha = 0.5$ for k_3, and k_4, and $\sigma = 21$ for Gaussian kernel.

(b) Results of GPR with the same tailored characteristic kernels are used in $\epsilon - SVR$ and Gaussian kernel.

Fig. 5: The results of ϵ-SVR and GPR with characteristic kernels k_1, k_2, k_3, k_4, and Gaussian kernel on SARCOS dataset.

3. Gaussian kernel with $\sigma = 0.05 \Rightarrow$ MSE $= 0.005$

Unexpectedly, Gaussian kernel worked better than justified kernels on the SO(3) group.

4.4 Abelian Semigroups: Classification of Human Actions

One example of Abelian semigroups are histograms. As many authors in computer vision area are working with kernel-based methods and histograms (for example, see the recent VOC2006 object classification challenge), it is worth studying kernel classes suitable for histogram-based information. We use the action descriptors introduced by Danafar and Gheissari [3], which are histograms of optical flow and capture both local and global information about actions. These feature vectors are described in Figure 6. We use the challenging human action video database of KTH [17]. It contains 6 types of human actions: walking, jogging, running, boxing, hand waving, and hand clapping, performed by 25 people in four different scenarios: outdoors (s1), outdoors with scale variations (s2), outdoors with different clothes (s3), and indoors (s4). Some samples from this dataset are shown in Figure 7.

Our action recognition approach is based on SVM. The database is divided into three parts: training, testing and validation. 8 subjects were used for training, 9 for test and 8 for validation. The validation data is first used to tune the hyper-parameter β of GHI kernel and our defined characteristic kernel with a 5-fold leave-one-out cross validation procedure. Danafar and Gheissari [3] recognized actions with SVMs and the GHI Kernel.

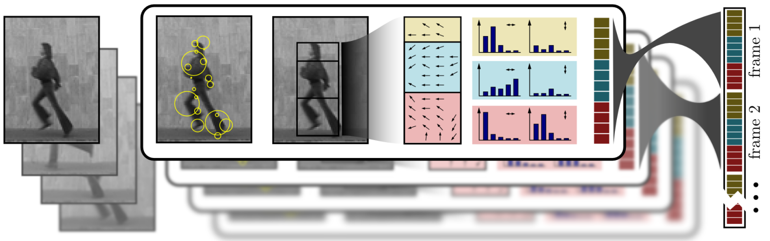


Fig. 6: The features used in our supervised classifier are described in [3]; a single feature vector (right) is computed for each sequence by concatenating data coming from each frame of the sequence (left). In each frame, Harris interest points are used to recover a tight bounding box, which is vertically partitioned in three regions. The topmost 1/5 of the bounding box approximately contains the head and the neck. The middle 2/5 contains the torso and hands, and the bottom 2/5 of the bounding box contains the legs. Such segmentation is obviously approximated, and the resulting features would still be usable in cases where the assumptions are not met. Flow data in each region is summarized in separate histograms for the horizontal and vertical directions.

The crucial condition an SVM kernel should satisfy is to be positive definite, meaning that the SVM problem is convex, and hence the solution of its objective function is unique. Characteristic kernels have positive definite property and have been shown to be more discriminative; because characterization can capture effective statistical and discriminative relationship between response variables from an explanatory variables [5], [6]. Our reported accuracy of 93.1% obtained with characteristic kernels is a very significant improvement with respect to the accuracy of 83.5% reported in [3], obtained using Histogram Intersection Kernel in the same setting. We also compared our characteristic kernel for histogram-based data to the Gaussian kernel, which is also characteristic but is not tailored to histogram-based data. In our experiments, the accuracy of Gaussian kernel is 33.8% which is much lower than our result of 93.1%. Confusion matrices in the three cases are reported in Figure 8. Therefore we conclude that our experimental results are due to our kernel being both characteristic and suitable for histogram-based data; removing any of the two properties results in a significant performance loss.

In Table 3 recognition results of various methods on the KTH dataset are compared. Our overall rate exceeds previously reported results and is comparable to 93.4% reported rate in [11], demonstrating superiority of our method. In [15] and [11], the authors benefited from stronger feature vectors as combination of shape and motion and reported high accuracy rates rather than motion feature which is used here. This concludes that the achievement of higher recognition rate with stronger histogram-based feature vector is promising.

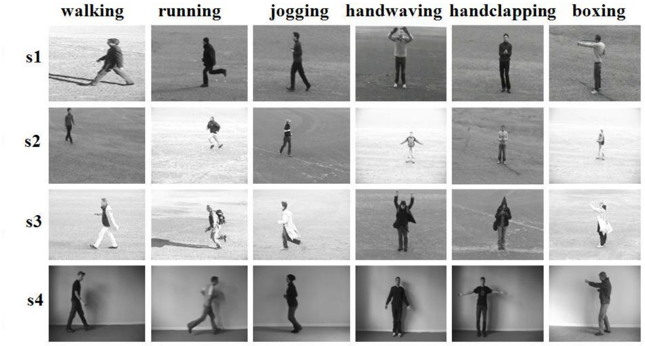


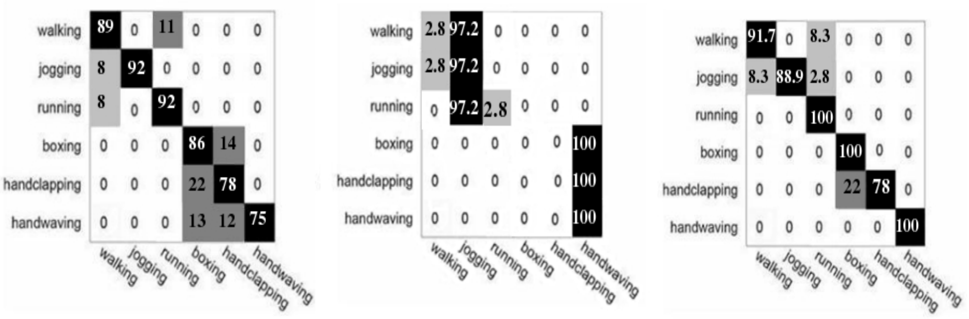
Fig. 7: Example images from video sequences in KTH dataset.

Table 3: Recognition results of various methods on the KTH dataset. The recognition rate reported by Jhuang et al. (2007) is obtained on video sequences from scenarios 1 and 4 only. Other reported rates are on all scenarios.

Method	Recognition rate %
SVM by charac. Kernel	93.1
Lin et al. [11]	93.4
Schindler and Van Gool [15]	92.7
Jhuang et al. [10]	91.7
Danafar & Gheissari [3]	85.3
Niebles et al. [12]	83.3
Schüldt et al. [17]	71.7

5 Conclusion

We studied empirically characteristic kernels on structured domains, yielding powerful kernel-based methods for structured data. Characteristic kernels for periodic domains and SO(3) were applied to homogeneity testing, forward kinematics, forward dynamics, and inverse dynamics for robotics. Our methods outperformed other published methods on the 8-input Kin forward kinematics data set, and the 8-input Pumadyn forward dynamics data set. We also used tailored characteristic kernels on histogram-based action descriptors to recognize human actions in video. Our results on the KTH database of human actions are comparable to or better than those of previous state-of-the-art methods. Ongoing work aims at improving inverse dynamics results through multi-task kernel-based learning with our tailored characteristic kernels.



	walking	jogging	running	boxing	handclapping	handwaving
walking	89	0	11	0	0	0
jogging	8	92	0	0	0	0
running	8	0	92	0	0	0
boxing	0	0	0	86	14	0
handclapping	0	0	0	22	78	0
handwaving	0	0	0	13	12	75

(a) Results of GHI kernel as a positive definite kernel with $\beta = 1$ and overall accuracy rate of 85.3%.

	walking	jogging	running	boxing	handclapping	handwaving
walking	2.8	97.2	0	0	0	0
jogging	2.8	97.2	0	0	0	0
running	0	97.2	2.8	0	0	0
boxing	0	0	0	0	0	100
handclapping	0	0	0	0	0	100
handwaving	0	0	0	0	0	100

(b) Results of Gaussian kernel as a characteristic kernel with $\sigma = 21$ and overall accuracy rate of 33.8%.

	walking	jogging	running	boxing	handclapping	handwaving
walking	91.7	0	8.3	0	0	0
jogging	8.3	88.9	2.8	0	0	0
running	0	0	100	0	0	0
boxing	0	0	0	100	0	0
handclapping	0	0	0	22	78	0
handwaving	0	0	0	0	0	100

(c)Results of the tailored characteristic kernel on histogram-based data with $\beta = 0.001$ and overall accuracy rate of 93.1%.

Fig. 8: Confusion matrices obtained on the KTH dataset with descriptors [3], using SVM and the indicated kernels. Figure(a) shows the recognition rates of histogram Intersection kernel which is a positive definite but not a characteristic kernel. Figure (b) denotes the result of a characteristic kernel (Gaussian) which is not tailored to histogram-based information. Figure (c) is the result of characteristic kernel which is tailored to histograms.

References

1. Boughorbel, S., Tarel, J.P., Boujemaa, N.: Generalized Histogram Intersection Kernel for image recognition. In IEEE International Conference on Image Processing, vol.3, pp.161-4 (2005).
2. Craig, J.I.: Introduction to Robotics, mechanics and control. Prentice Hall, 3rd edition (2004).
3. Danafar, S., and Gheissari, N.: Action recognition for surveillance application using optic flow and SVM. In Asian Conference on Computer Vision, vol.2, pp.457-466 (2007).
4. Fisher, N.I.: Statistical analysis of circular data, Cambridge University Press (1993).
5. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel Measures of Conditional Dependence. In Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference, pp.489-496. (Eds.) Platt, J. C., D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).
6. Fukumizu, K., Bach, F.R., and Jordan, M.I.: Dimensionality reduction for supervised learning with reproducing kernel Hilbert Spaces, JMLR, vol. 5, pp.73-99 (2004).
7. Fukumizu, K., Sriperumbudur B.K., Gretton, A., and Schölkopf, B.: Characteristic kernels on groups and semigroups, In Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference, pp. 473-480. (Eds.) Koller, D., D. Schuurmans, Y. Bengio, L. Bottou, Curran, Red Hook, NY, USA (06 2009)
8. Gretton, A., Borgwadt, B.K., Rasch, M., Schölkopf, B., and Smola, A.: In Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, pp.513-520. (Eds.) Schölkopf, B., J. Platt, T. Hofmann, MIT Press, Cambridge, MA, USA (2007).

9. Gretton, A., Fukumizu, K., Teo C.H., Song, L., Schölkopf, B., and Smola, A.: A Kernel Statistical Test of Independence. In Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference, pp.585-592. (Eds.) Platt, J. C.,D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).
10. Jhuang H., Serre T., Wolf, L., and Poggio, T.: A biologically inspired system for action recognition. In Internation Conference on Computer Vision,pp.1-8 (2007).
11. Lin, Z., Jiang, Z., and Davis, L.S.: Recognizing Actions by Shape-Motion Prototype Trees. In IEEE International Conference on Computer Vision (ICCV 2009), Kyoto, Japan.
12. Niebles, J.C., Wang, H., and Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In IJCV, Vol. 79, No. 3, pp. 299-318 (2008).
13. Nguyen-Tuong, D., Peters, J., Seeger, M., and Schölkopf, B.: Learning Inverse Dynamics: a Comparison. Advances in Computational Intelligence and Learning: Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2008), pp.13-18. (Eds.) Verleysen, M. d-side, Evere, Belgium (2008).
14. Rasmussen, C.E., and Williams, K.A.: Gaussian processes for machine learning. MIT-press, Massachusetts Institute of Technology (2006).
15. Schindler, K., and Van Gool, L.: Action snippets: How many frames does human action recognition require?, In International conference on Computer Vison and Pattern Recognition (CVPR), pp.1-8 (2008).
16. Schölkopf, B., and Smola, A.: Learning with kernels. MIT press, Cambridge, MA (2002).
17. Schüldt, C., Laptev, I., and Caputo, B.: Recognizing human actions: a local SVM approach. Proceedings of the 17th International Conference on In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, Vol. 3, pp. 32-36 (2004).
18. Smola, A.J., and Schölkopf, B.: A tutorial on Support Vector Regression.: NeuroCOLT Technical Report TR-98-030 (1998).
19. Smola, A.J., Gretton, A., Song, L., Schölkopf, B.: A Hilbert Space Embedding for Distributions. In Algorithmic Learning Theory: 18th International Conference (ALT 2007), 13-31. (Eds.) Hutter, M., R. A. Servedio, E. Takimoto, Springer, Berlin, Germany (2007).
20. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Lanckeriet, G.R.G., and Schölkopf, B.: Injective Hilbert space embeddings of probability measures.Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008), pp.111-122. (Eds.) Servedio, R. A., T. Zhang, Omnipress, Madison, WI, USA (2008).
21. Yeung, D.Y., and Zhang, Yu.: Learning inverse dynamics by Gaussian Process Regression under the multi-task learning framework. In the Path to Autonomous Robots, G.S.Sukatme (ed.), pp.131-142, Springer (2009).
22. Waterhouse, S.: PhD Thesis on Mixtures of Experts Available (1998).