

# Sparse Kernel Machines for Discontinuous Registration and Nonstationary Regularization

Christoph Jud, Nadia Möri, Philippe C. Cattin

Dept. of Biomedical Engineering, University of Basel, Switzerland

christoph.jud@unibas.ch

## Abstract

*We present a novel approach where we address image registration with the concept of a sparse kernel machine. We formulate the registration problem as a regularized minimization functional where a reproducing kernel Hilbert space is used as transformation model. The regularization comprises a sparsity inducing  $l_1$ -type norm and a well known  $l_2$  norm. We prove a representer theorem for this type of functional to guarantee a finite dimensional solution. The presented method brings the advantage of flexibly defining the admissible transformations by choosing a positive definite kernel jointly with an efficient sparse representation of the solution. As such, we introduce a new type of kernel function, which enables discontinuities in the transformation and simultaneously has nice interpolation properties. In addition, location-dependent smoothness is achieved within the same framework to further improve registration results. Finally, we make use of an adaptive grid refinement scheme to optimize on multiple scales and for a finer control point grid at locations of high gradients. We evaluate our new method with a public thoracic 4DCT dataset.*

## 1. Introduction

Non-rigid image registration is a central problem in many medical image analysis tasks. The aim of image registration is, to align two similar images in a way, that a target image can be expressed through a reference image by a spatial transform mapping. To recover meaningful anatomical changes, smooth transformations are often desirable. In contrast to that, in abdominal imaging, sliding organ boundaries require discontinuous transforms for accurate alignments, which imposes challenges on defining proper transformation models.

In this paper, we formulate image registration as a combined  $l_1$ -type and  $l_2$  regularized minimization problem,

whose regularization favors sparse solutions. We define the space of admissible transform mappings as an infinite dimensional reproducing kernel Hilbert space (RKHS) and prove the corresponding representer theorem in order to guarantee a finite solution. The theorem states that a minimizer of the discretized functional lies within a finite dimensional linear subspace of the infinite dimensional RKHS, and this subspace is spanned by the basis functions placed at the spatially sampled finite number of points. To the best of our knowledge, this has not been proven for this kind of  $l_1$ -type regularized functionals, so far. The application of the functional to image registration is new as well. It has the advantage that the desired properties of the resulting transform mapping can be specified directly by a positive definite kernel function. To cope with cases, where discontinuities in the transformation are desirable, we introduce a new compactly supported kernel function that allows for such discontinuous transforms and simultaneously has nice interpolation properties. To demonstrate the flexibility of our approach we further introduce a nonstationary kernel function that yields smoother transformations at locations with bony structures and less smooth ones otherwise. For the optimization, we adopt an adaptive grid approach, where the control point grid is only refined where the parameters' gradient magnitude is non-zero. We evaluate our method on the publicly available 4DCT dataset of the POPI model [21], where we achieve a state-of-the-art registration performance without using manually defined image masks.

The parametric way to approach non-rigid image registration has been extensively studied for more than two decades [11, 12]. Based on the introduction of the free-form deformations (FFD) into registration [12], a lot of advanced registration approaches have been successfully applied to medical images. For a comprehensive overview about non-rigid image registration methods, including non-parametric and discrete approaches, we refer to the survey [19]. Lately, discontinuity preserving attempts appeared increasingly [1, 3, 5, 10, 14]. In [18], FFD have been applied along with  $l_1$  regularization on the transform parameters in

multiple scales jointly. In [22], total variation regularization on the displacement field has been approximated.  $l_1$  regularization causes the solution to be sparse. However, for the continuous objective and the underlying infinite dimensional function space, there is no finite solution in general. Because of the parametrization therefore, an infinite dimensional part of the solution space is lost.

$l_1$  regularization has a long tradition in machine learning [13, 20] and specifically in kernel methods [24]. Despite their flexibility and profound theory, kernel methods are rarely used for registration [7, 9, 16]. The flexibility originates from the fact that the properties of the admissible transformations can be modeled directly through a positive definite kernel function. This stands in contrast to FFD-like approaches, where *standard* regularizers specify the transformation properties indirectly through differential operators and hence are conceptually rather rigid. Furthermore, as stated above, with the help of reproducing kernel Hilbert spaces, finite solutions are guaranteed by representer theorems [2].

In the following, we first define the transformation model and the type of functional we consider for image registration. This is followed by concrete models where we introduce kernels which lead to discontinuous transformations and spatially varying smoothness properties.

## 2. Method

Let  $I_R, I_T : \mathcal{X} \rightarrow \mathbb{R}$  be a reference and a target image, which map the  $d$ -dimensional input domain  $\mathcal{X} \subset \mathbb{R}^d$  to the real numbers i.e. image intensities. Furthermore, let us define a spatial transform mapping  $f : \mathcal{X} \rightarrow \mathbb{R}^d$ . Image registration can be formulated as a regularized functional minimization problem

$$\min_f \mathcal{J}[f], \quad \mathcal{J}[f] := \mathcal{D}[I_R, I_T, f] + \eta \mathcal{R}[f], \quad (1)$$

where  $\mathcal{D}$  is a dissimilarity measure between the transformed reference image and the target image. In this paper, we focus on measures which integrate over a loss function  $\mathcal{L} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ :

$$\mathcal{D}[I_R, I_T, f] := \int_{\mathcal{X}} \mathcal{L}(I_R(x + f(x)), I_T(x)) dx. \quad (2)$$

which could be e.g. the squared loss  $(s - s')^2$ . The regularizer  $\mathcal{R}$  ensures certain properties of the transformations.

As transformation model, we define  $f$  through a reproducing kernel Hilbert space (RKHS)

$$\mathcal{H} := \left\{ f \left| f(x) = \sum_{i=1}^{\infty} c_i k(x_i, x), \right. \right. \\ \left. \left. x_i \in \mathcal{X}, \quad c_i \in \mathbb{R}^d, \quad \|f\| < \infty \right. \right\}, \quad (3)$$

where  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a positive definite kernel and induces  $\mathcal{H}$ , cf. a comprehensive introduction to kernel methods and RKHS in [4]. As regularizer, the RKHS norm

$$\mathcal{R}[f] := \|f\| = \sqrt{\langle f, f \rangle}, \quad (4)$$

with the inner product  $\langle \cdot, \cdot \rangle$  of  $\mathcal{H}$ , is usually defined which measures the magnitude of  $f$ . This is no coincidence because using exactly (4) as regularizer, the *standard* representer theorem (see Section 2.2) allows for the discretization of the original objective (1) without loosing the guarantee of a finite dimensional minimizer. However, we are additionally interested in a sparsity inducing norm as a regularization, which we will define next.

### 2.1. Definition of an $l_1$ -type Norm

Without loss of generality, we focus on  $d = 1$  to simplify notation. All findings can be generalized to arbitrary dimensions  $d \in \mathbb{N}_+$ . Let  $\mathcal{X}$  be the sample space where  $x_i \in \mathcal{X}$  and  $\mathcal{H}$  be an RKHS on  $\mathcal{X}$  induced by the *strictly* positive definite kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ . Consider the subset  $\mathcal{H}_0 \subset \mathcal{H}$  for a set  $\{x_i\}_{i=1}^N$  of pairwise distinct samples

$$\mathcal{H}_0 = \left\{ f_0 \in \mathcal{H} \left| f_0(\cdot) = \sum_{i=1}^N c_i k(x_i, \cdot), c_i \in \mathbb{R} \right. \right\}. \quad (5)$$

$\mathcal{H}_0$  is a finite dimensional linear subspace of  $\mathcal{H}$ . Since  $k$  is a positive definite kernel, the matrix  $K = (k(x_i, x_j))_{i,j}$  is positive definite and therefore  $\{k(\cdot, x_j)\}_{j=1}^N$  forms a basis of  $\mathcal{H}_0$  which can be orthogonalized to a basis  $\{\psi_j(x)\}_{j=1}^N$  w.r.t.  $\langle \cdot, \cdot \rangle$ . Hence, there exist  $\{\lambda_k\}_{k=1}^N \subset \mathbb{R}$  such that we can write

$$k(\cdot, x_j) = \sum_{k=1}^N \lambda_k \psi_k(\cdot), \quad j = 1, \dots, N. \quad (6)$$

Let us define a projection  $P : \mathcal{H} \rightarrow \mathcal{H}_0$  onto the subspace  $\mathcal{H}_0$  by

$$P(f) := \sum_{i=1}^N \langle f, \psi_i \rangle \psi_i. \quad (7)$$

$P$  is well-defined, since  $\langle \cdot, \cdot \rangle$  is well-defined. It holds that  $P(f) = f$  for all  $f \in \mathcal{H}_0$  and  $P^2 = P$ . Now, every  $f \in \mathcal{H}$  can be decomposed by  $f = P(f) + v$ . It holds that

$$\langle v, k(\cdot, x_i) \rangle = 0, \quad i = 1, \dots, N, \quad (8)$$

i.e.  $v \in \mathcal{H}_0^\perp$ . To prove this, observe that for  $k = 1, \dots, N$ :

$$\begin{aligned} \langle P(f), \psi_k \rangle &= \left\langle \sum_{i=1}^N \langle f, \psi_i \rangle \psi_i, \psi_k \right\rangle \\ &= \sum_{i=1}^N \langle f, \psi_i \rangle \langle \psi_i, \psi_k \rangle = \langle f, \psi_k \rangle. \end{aligned}$$

This implies that

$$\langle v, \psi_k \rangle = \langle f - P(f), \psi_k \rangle = \langle f, \psi_k \rangle - \langle P(f), \psi_k \rangle = 0.$$

Hence:

$$\langle v, k(\cdot, x_i) \rangle = \left\langle v, \sum_{k=1}^N \lambda_k \psi_k \right\rangle = \sum_{k=1}^N \lambda_k \langle v, \psi_k \rangle = 0,$$

for  $i = 1, \dots, N$  and thus (8) is proven.

We therefore conclude that for each  $f \in \mathcal{H}$  there are unique  $c_i$  such that

$$f(x) = \sum_{i=1}^N c_i k(x_i, x) + v, \quad (9)$$

where  $v$  is orthogonal to  $\mathcal{H}_0$ . Keeping that in mind, we define the norm

$$\|f\|_1 := \sum_{i=1}^N |c_i| + \|v\|, \quad (10)$$

where  $f \in \mathcal{H}$  and decomposed as described above. Since  $\|\cdot\|$  is well-defined and the  $c_i$  are unique,  $\|\cdot\|_1$  is well-defined. As  $\|\cdot\|$  and  $|\cdot|$  are norms,  $\|\cdot\|_1$  is a norm as well.

## 2.2. A Representer Theorem

**Theorem 1.** *Let the training data  $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathbb{R} \mid i = 1, \dots, N\}$ , a loss function  $\mathcal{L} : \mathcal{X} \times \mathbb{R} \times \mathcal{H} \rightarrow \mathbb{R} \cup \{\infty\}$  and two functions  $g : \mathbb{R} \rightarrow \mathbb{R}$  and  $h : \mathbb{R} \rightarrow \mathbb{R}$  be given. If one of the two functions  $g$  or  $h$  is strictly increasing and the other one is nondecreasing, the minimization problem*

$$\min_{f \in \mathcal{H}} \sum_{i=1}^N \mathcal{L}(x_i, y_i, f(x_i)) + g(\|f\|_{\mathcal{H}}) + h(\|f\|_1) \quad (11)$$

has a minimizer taking the form

$$f(x) = \sum_{i=1}^N c_i k(x_i, x), c_i \in \mathbb{R}. \quad (12)$$

In particular, we are interested in the case where  $g(x) = 0$ ,  $h(x) = x$  (i.e.  $g$  nondecreasing and  $h$  strictly increasing) and  $\mathcal{L}(x_i, y_i, f(x_i)) = \mathcal{L}(I_R(x_i + f(x_i)), I_T(x_i))$ . First, we consider the case where  $h(x) = 0$  and follow the argumentation of [15]. We call this part *standard* representer theorem. This is followed by the proof for the full theorem.

*Proof.* We define a map from  $\mathcal{X}$  into the space of functions, mapping  $\mathcal{X}$  into  $\mathbb{R}$ , denoted as  $\mathbb{R}^{\mathcal{X}}$  via

$$\phi : \mathcal{X} \rightarrow \mathbb{R}^{\mathcal{X}}, \quad x \rightarrow k(\cdot, x). \quad (13)$$

Since  $k$  is a reproducing kernel, evaluation of the function  $\phi(x)$  on the point  $x'$  yields

$$(\phi(x))(x') = k(x', x) = \langle \phi(x'), \phi(x) \rangle \quad \forall x, x' \in \mathcal{X}. \quad (14)$$

As described above in Section 2.1, we decompose  $f \in \mathcal{H}$  as  $f = \sum_{i=1}^N c_i \phi(x_i) + v$  for unique  $c \in \mathbb{R}^N$  and  $v \in \mathcal{H}$  satisfying  $\langle v, \phi(x_j) \rangle = 0$ ,  $j = 1, \dots, N$ . Using (14), the application of  $f$  to an arbitrary training point  $x_j$  yields

$$f(x_j) = \left\langle \sum_{i=1}^N c_i \phi(x_i) + v, \phi(x_j) \right\rangle = \sum_{i=1}^N c_i \langle \phi(x_i), \phi(x_j) \rangle,$$

which is independent of  $v$ . Consequently, the loss function term of (11) is independent of  $v$ .

**Standard Representer Theorem** Let  $h(x) = 0$  and  $g$  be strictly increasing. As for the second term in (11), since  $v$  is orthogonal to  $\sum_{i=1}^N c_i \phi(x_i)$ , and  $g$  is strictly increasing, we get

$$\begin{aligned} g(\|f\|) &= g\left(\left\|\sum_i c_i \phi(x_i) + v\right\|\right) \\ &= g\left(\sqrt{\left\|\sum_i c_i \phi(x_i)\right\|^2 + \|v\|^2}\right) \\ &\geq g\left(\left\|\sum_i c_i \phi(x_i)\right\|\right), \end{aligned} \quad (15)$$

with equality occurring iff  $v = 0$ . Setting  $v = 0$  thus affects neither the first nor the third term of (11), while strictly reducing the second term. Hence, any minimizer must have  $v = 0$ . Consequently, any solution takes the form  $f = \sum_{i=1}^N c_i \phi(x_i)$ , that is, using (14)

$$f(\cdot) = \sum_{i=1}^N c_i k(\cdot, x_i). \quad (16)$$

**Generalized Representer Theorem** Let's look at the case where  $g$  is nondecreasing and  $h$  is strictly increasing. The second case can be proved with exactly the same argument, just by switching  $g$  and  $h$ . Then, (15) is still valid. However, equality can also occur when  $v \neq 0$  since  $g$  is nondecreasing. For  $h$  we observe that

$$h(\|f\|_1) = h\left(\sum_{i=1}^N |c_i| + \|v\|\right) \geq h\left(\sum_{i=1}^N |c_i|\right), \quad (17)$$

with equality iff  $\|v\| = 0$ . Note that  $\|f\|_1 = 0 \Leftrightarrow \|f\| = 0$ . Now, we conclude from Equations (17) and (15) that  $g(\|f\|) + h(\|f\|_1)$  can be minimized iff  $v = 0$ , which means that any minimizer of (11) lives in  $\mathcal{H}_0$  and the theorem is thus proven.  $\square$

### 2.3. Compactly Supported Kernels

We have derived the framework in which we can choose the positive definite kernel  $k$  to complete the transformation model. A kernel which yields smooth interpolants is e.g. the *Gaussian* kernel  $k_g(x, y) = \exp(-\|x - y\|^2/\sigma_g^2)$ ,  $\sigma_g > 0$ . Since  $k_g$  has infinite support, however, the computational complexity of evaluating the objective (1) is  $\mathcal{O}(N^2)$  kernel evaluations. Especially when dealing with large images, it is therefore important that the applied kernel has compact support to reduce the number of summands in  $f$ . A kernel which has compact support and simultaneously yields interpolants in  $C^4$  is e.g. a  $C^4$  *Wendland* kernel [23]

$$k_4(x, y) = \psi_{3,2}\left(\frac{\|x - y\|}{\sigma_4}\right), \quad (18)$$

$$\psi_{3,2}(r) = (1 - r)_+^6 \frac{3 + 18r + 35r^2}{1680},$$

with  $\sigma_4 > 0$ ,  $d \leq 3$ ,  $a_+ = \max(0, a)$  and  $\psi_{3,2}$  being the *Wendland* function of the second kind and positive definite in three dimensions.

Since  $k_4$  yields basis functions in  $C^4$  and  $C^4$  functions are closed with respect to linear combinations no discontinuities can be represented in  $f$ . However, to still allow for discontinuous transformations, we multiplicatively superimpose a  $C^0$  *Wendland* kernel

$$k_c(x, y) = \psi_{3,0}\left(\frac{\|x - y\|}{\sigma_0}\right), \quad (19)$$

$$\psi_{3,0}(r) = (1 - r)_+^2,$$

with  $\sigma_0 > 0$ ,  $d \leq 3$  and  $\psi_{3,0}$  being the *Wendland* function of the zeroth kind and positive definite in three dimensions. This kernel has a removable discontinuity at  $x = y$  and hence enables discontinuous transforms  $f$ . Because the multiplication of two valid kernels is again a valid kernel function (cf. [17]), we profit from the properties of both kernels and define

$$k_c(x, y) = k_4(x, y)k_0(x, y).$$

The maximum support range of  $k_c$  is  $\min(\sigma_4, \sigma_0)$ . In Figure 1, basis functions of  $k_4$ ,  $k_0$  and  $k_c$  with different values of  $\sigma_4$  and  $\sigma_0$  for  $k_4$  and  $k_0$  are plotted. With increasing  $\sigma_0$ , the combined kernel becomes similar to the  $C^4$  *Wendland* kernel and if  $\sigma_4$  is large compared to  $\sigma_0$ ,  $k_c$  becomes similar to the  $C^0$  *Wendland* kernel.

**Spatially Varying Kernels** There is a great interest in inhomogeneous transformation models, since tissues vary among the spatial domain. Moreover, we think it is crucial that the degree of smoothness can be influenced by the user. Otherwise, an improper result has to be discarded which

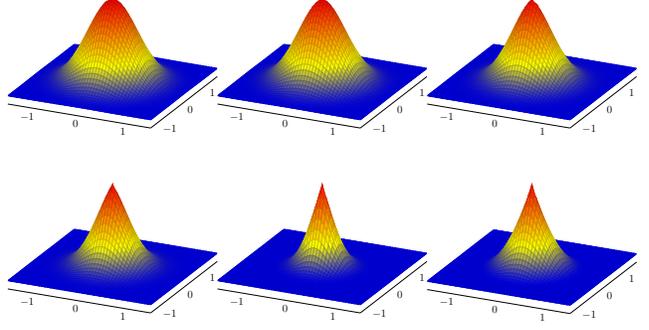


Figure 1: Samples of 2d basis functions at zero. For all samples  $\sigma_4 = 2$ . First row:  $k_4$ ,  $k_{c,\sigma_0=16}$ ,  $k_{c,\sigma_0=4}$ . Second row:  $k_{c,\sigma_0=1.5}$ ,  $k_{c,\sigma_0=1}$ ,  $k_{0,\sigma_0=1}$ .

could have been improved with inhomogeneous smoothness. In [8], it is shown how to obtain spatially varying kernels from stationary isotropic kernels. We will apply this technique to our combined kernel function.

Let  $\Sigma(x)$  be a  $d \times d$  matrix-valued function which is positive definite for all  $x \in \mathcal{X}$ , ( $\Sigma_i := \Sigma(x_i)$ ) and let us define the quadratic form

$$Q_{ij} = (x_i - x_j)^T ((\Sigma_i + \Sigma_j)/2)^{-1} (x_i - x_j). \quad (20)$$

Then, for a stationary isotropic kernel  $k_s$ ,

$$k_{ns}(x_i, x_j) = |\Sigma_i|^{1/4} |\Sigma_j|^{1/4} \left| \frac{\Sigma_i + \Sigma_j}{2} \right|^{-1/2} k_s(\sqrt{Q_{ij}}) \quad (21)$$

is a valid positive definite *nonstationary* kernel, which was shown in [8]. For the  $C^4$  *Wendland* kernel  $k_4$  we only consider diagonal matrices  $\Sigma(x) := \sigma_x I_{d \times d}$ ,  $\sigma_x > 0$  and define

$$k_4^{ns}(x, y) = \Psi_{\sigma_x, \sigma_y} \psi_{3,2}\left(\Phi_{\sigma_x, \sigma_y} \frac{\|x - y\|}{\sigma_4}\right), \quad (22)$$

where

$$\Psi_{\sigma_x, \sigma_y} = |\sigma_x|^{1/4} |\sigma_y|^{1/4} \left| \frac{\sigma_x + \sigma_y}{2} \right|^{-1/2},$$

$$\Phi_{\sigma_x, \sigma_y} = \sqrt{\frac{2}{\sigma_x + \sigma_y}}$$

and  $\sigma_x, \sigma_y$  act as *weights* on the smoothness. Accordingly, we define the *nonstationary*  $C^0$  *Wendland* kernel

$$k_0^{ns}(x, y) = \Psi_{\sigma_x, \sigma_y} \psi_{3,0}\left(\Phi_{\sigma_x, \sigma_y} \frac{\|x - y\|}{\sigma_0}\right)$$

and  $k_c^{ns} = k_4^{ns} k_0^{ns}$ . To control the local amount of smoothness, we finally define a weight image  $\sigma : \mathcal{X} \rightarrow \mathbb{R}_{>0}$ . Larger values of  $\sigma(x)$  result in transformations  $f$  which are locally smoother.

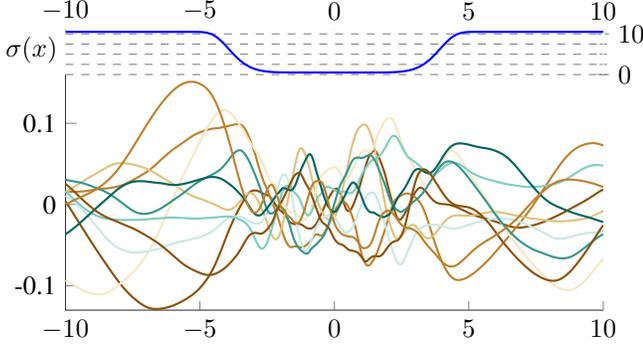


Figure 2: Random samples of admissible 1d transformations  $f$  assuming a zero-mean Gaussian process with covariance function  $k_4^{ns}$  with  $\sigma_4 = 3$  and a 1d weight image  $\sigma(x) = 10.5 - 10e^{-\left(\frac{|x|}{4}\right)^8}$ .

In Figure 2, we show one-dimensional random samples of  $f$  using the nonstationary  $C^4$  Wendland kernel. By reducing the value  $\sigma(x)$  in the middle of the interval, the smoothness locally decreases.

Using the weight image  $\sigma$ , prior knowledge about local smoothness of the transformation can be incorporated. In our experiments (see Section 3), we will construct weight images, where strong smoothness is enforced for bony structures and less smoothness otherwise.

## 2.4. Optimization

Since the registration functional (1) is non-convex, we locally optimize (1) using a first order variant of the projected subgradient method of [13] (Example 11.8), which can handle the non-differentiable regularizer.

**Two-Metric Subgradient Projection Iteration** Consider the objective

$$\min_{c \in \mathbb{R}^{dN}} L(c) + \sum_{i=1}^N \lambda_i |c_i|, \quad (23)$$

where  $L$  is differentiable and corresponds to our dissimilarity measure, which depends on the parameters  $c$  and the scalars  $\lambda_i \geq 0$ . Note that we set all  $\lambda_i$  to the regularization weight  $\eta$ . The minimum-norm subgradient  $z_i^k$  at the  $k$ th iteration is computed as

$$z_i^k = \begin{cases} \nabla_i L(c) + \lambda_i \operatorname{sgn}(c_i), & |c_i| > 0 \\ \nabla_i L(c) + \lambda_i, & c_i = 0, \nabla_i L(c) < -\lambda_i \\ \nabla_i L(c) - \lambda_i, & c_i = 0, \nabla_i L(c) > \lambda_i \\ 0, & c_i = 0, -\lambda_i \leq \nabla_i L(c) \leq \lambda_i. \end{cases} \quad (24)$$

We define the two-metric projection step as

$$c^{k+1} = \mathcal{P}_{\mathcal{O}}[c^k - \alpha z^k, c^k], \quad (25)$$

where the orthant projection is defined as

$$\mathcal{P}_{\mathcal{O}}(y, x)_i = \begin{cases} 0, & \text{if } x_i y_i < 0 \\ y_i, & \text{otherwise.} \end{cases} \quad (26)$$

The parameter  $\alpha$  is the step-size for the current iteration. We adjust  $\alpha$  in an adaptive manner [6], while we approximate  $L$  by uniformly sampling  $n$  image points  $\{x_i\}_i^n \subset \Omega$ .

**Multiple Scale Optimization** To reduce local optima, we successively approximate the transform mapping  $f$  with decreasingly scaled transforms  $f \approx \tilde{f}_0 + \tilde{f}_1 + \dots + \tilde{f}_l$ . In our case, we decrease the width parameters  $\sigma_4, \sigma_0$  of the kernels in each level. Thus, the transformations tend to be smoother in lower levels than in the succeeding ones. Note, since we perform stochastic sampling of  $L$  we do not scale the images.

To reduce the computational demand, we sample *control points* on a uniform grid and place the basis functions onto these grid points. Furthermore, each control point can be enabled or disabled similar to the non-uniform grid approach [11]. We define the refinement measure  $\xi(x_i) = \left| \frac{\partial}{\partial c_i} \mathcal{J}[f] \right|$ , where  $c_i$  is only enabled if  $\xi(x_i)$  exceeds a threshold  $\tau$ . As we perform gradient based optimization, we assume that the registration cannot be improved significantly where  $\xi(x_i)$  is small, since exactly the derivative of  $\mathcal{J}[f]$  is considered to proceed a decent step. This is certainly not always true.

## 3. Results

We evaluate our new Sparse Kernel Machine (SKM) on a synthetic example to first investigate sliding boundaries. Then we register a publicly available 4DCT dataset, where ground truth landmarks are available. The challenges here are mainly the sliding organ boundaries, e.g. the pleural cavity between the lung and the vertebrae.

For all experiments, we used the squared loss function and exclusively performed  $l_1$ -type regularization. Furthermore, we have scaled  $k_4$  by 560 such that  $k_4(x, x) = 1, \forall x \in \mathcal{X}$  and we set the threshold  $\tau = 0$ .

### 3.1. Synthetic Example

For this experiment, we have been inspired by the synthetic examples of [14]. The reference and target images are depicted in Figure 3. First, we experimentally confirm our previous statement that non-differentiable basis functions are superior to differentiable ones in representing discontinuous transformations. For that, we discretized the ground truth transformation  $f_{gt} \in \mathbb{R}^{N \times 2}$  and projected it into our parametric space  $\mathcal{H}_0$  by

$$c_{gt} = f_{gt} K^{-1}. \quad (27)$$

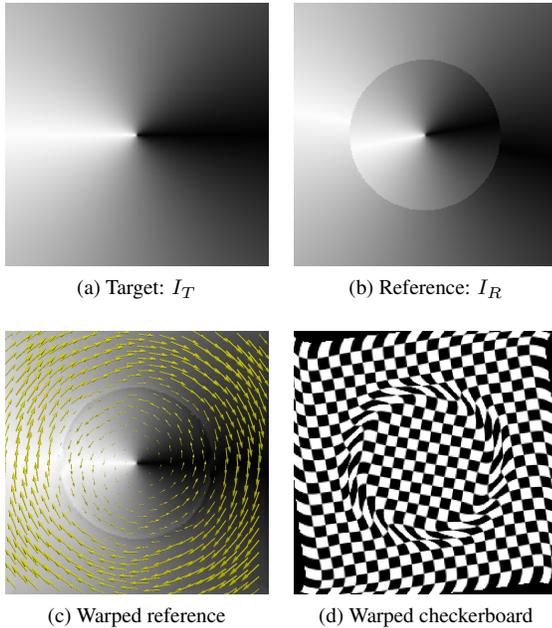


Figure 3: Synthetic example:  $I_T(x) = \frac{x \times e_1}{\pi}$ , where  $e_1$  is a unit vector and the origin is in the center of the image. In the reference image, the circular inner as well as the outer region have been rotated by 10 degrees but in opposite direction.

For comparison, we define the error  $\varepsilon := \frac{1}{N} \|f_{gt} - c_{gt}^T K\|$  and evaluate (27) on a resolution of  $120 \times 120$  px and pixel spacing  $0.3 \times 0.3$  mm<sup>2</sup>. The Gram matrix  $K$  was computed using the  $C^4$  Wendland kernel  $k_4$  with different  $\sigma_4$  (first curve) and the introduced combined and non-differentiable kernel  $k_c$  having fixed  $\sigma_4 = 30$  and with a varying  $\sigma_0$  (second curve), see Figure 4. With the upper values of the sig-

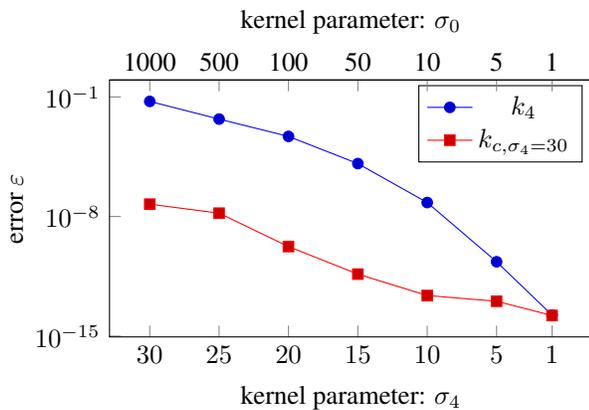


Figure 4: Comparison of projection error between the kernel  $k_4$  and the combined kernel  $k_c$  with fixed  $\sigma_4 = 30$ .

Table 1: Average TRE [mm] and maximum TRE (subscript) of the first 40 landmarks with respect to image 1.

#	No reg.	FFD	pTV	SKM	<i>ns.</i> SKM
0	0.48 2.4	0.79 1.5	0.72	0.67 1.6	0.66 1.5
2	0.49 2.6	0.81 2.2	0.71	0.65 1.7	0.65 1.7
3	2.19 6.6	1.14 2.8	1.12	1.17 2.5	1.17 2.9
4	4.33 10	1.11 2.4	1.01	1.05 2.1	1.07 2.3
5	5.75 12	1.11 3.2	1.11	1.11 3.1	1.13 3.1
6	6.01 14	1.20 3.2	1.03	0.99 2.7	1.00 2.6
7	5.03 12	1.20 3.0	1.06	1.08 3.2	1.05 3.1
8	3.68 6.2	0.88 2.3	0.84	0.76 1.6	0.75 1.7
9	2.07 4.5	0.92 2.0	0.81	0.83 2.1	0.83 2.1
$\emptyset$	3.35 14	1.02 3.2	0.93	0.92 3.2	0.92 3.1

mata, where the two kernels have almost the same shape, the average error for the combined kernel  $k_c$  was six orders of magnitude smaller, which indicates that it is better suited as transformation model for this kind of problem.

We additionally registered this example on a resolution of  $240 \times 240$  px and  $0.3 \times 0.3$  mm<sup>2</sup> pixel spacing using the combined kernel  $k_c$ . We optimized on four scales with  $\sigma_4 = \{100, 50, 25, 10\}$ ,  $\sigma_0 = \{12, 8, 4, 2\}$  and  $\eta = \{10^{-9}, 10^{-8}, 10^{-7}, 10^{-7}\}$ . The isotropic control point grid spacing was  $\{2, 1.3, 0.6, 0.3\}$ . The number of image samples  $n$  was set to the full image resolution. The inner and outer part have been accurately and smoothly matched, cf. Figure 3. However, there is room for improvement in a thin band near the discontinuous border. The average displacement error was 1.24 mm. To avoid local optima, a small step-size was needed. Hence, the convergence rate was rather weak.

### 3.2. POPI Model

The POPI model [21] is a temporal sequence of 10 thoracic CT images having dimensions  $482 \times 360 \times 141$  and pixel spacing  $0.97 \times 0.97 \times 2$  mm<sup>2</sup>. All the background has been removed by the publishers and set to  $-1000$ HU. The target registration error<sup>1</sup> (TRE) is computed using the 40 ground truth landmarks which are available as well. We compare our SKM with the classic FFD [12] where the transformations have been taken from the POPI homepage<sup>2</sup>. Additionally, we compare to a recent method pTV [22], where we have taken the TRE values of their publication (see Table 1). We manually optimized parameters for the image 6 and used the same configuration for all other time steps. The target image  $I_T$  is image 1. We optimized on three scales with  $\sigma_4 = \{64, 44, 32\}$ ,  $\sigma_0 = \{500, 100, 20\}$  and  $\eta = \{0.01, 0.005, 0.001\}$ . The isotropic control point

<sup>1</sup>Mean Euclidean distance between ground truth target landmarks and reference landmarks which have been warped by the resulting  $f$ .

<sup>2</sup><http://www.creatis.insa-lyon.fr/rio/popli-model>

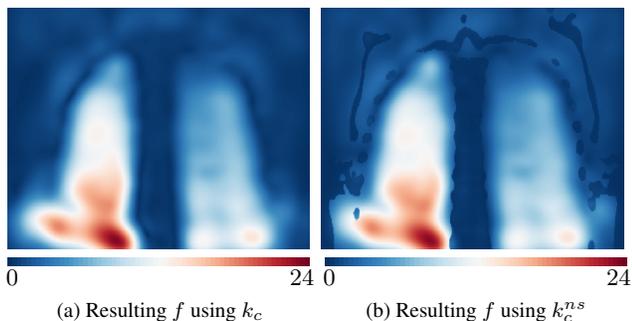


Figure 5: Magnitudes [mm] of the resulting  $f$  of a coronal slice for sample image 6. In (b), the effect of the weighting image  $\sigma$  can be clearly seen along the vertebrae, the ribs and the scapulae.

grid spacing was  $\{16, 11, 5\}$  resulting in  $\{24k, 74k, 775k\}$  parameters, where on average  $\{18\%, 22\%, 30\%\}$  of them were zero. The average number of sampled image points  $n$  was  $\{7k, 16k, 112k\}$ . We performed the experiments on two Intel Xeon E5 CPUs at 2.4GHz with an average processing time of  $\{0.2, 1.1, 10.7\}$  hours. Concerning the TRE, we achieved state-of-the-art results (see Table 1).

In a second run, we considered a weighting image for inhomogeneous regularization. The weighting has been generated considering the temporal intensity variance of a point and two threshold values indicating bone structure and lung tissue. For bone structure, the weight was set to 0.1 and to 1 otherwise. The image pairs were registered with the same parameters except for the regularization weight  $\eta = \{0.005, 0.001, 0.0004\}$ , resulting in  $\{21\%, 19\%, 28\%\}$  zero parameters. Although the derivation of the weighting image is rather ad hoc, it illustrates how the following problem can be tackled. Let us consider the registered image 6 in Figure 6. The lowest vertebra in the non-weighted method is teared down along with the lung and liver and is distorted in an unnatural way. With the different weighting of bony structures, one can see in the weighted version that the distortion could be reduced without an increase of the TRE.

## 4. Conclusion

In this paper, we presented a novel image registration framework where for the transformation model we made use of an infinite dimensional reproducing kernel Hilbert space. We defined an  $l_1$ -type norm and showed that it is well defined in this RKHS. We additionally proved a representer theorem for the derived functional in order to guarantee a finite dimensional solution. The new  $l_1$ -type norm induces sparsity in the transform parameters and hence yields an efficient representation of the found solution.

To target the alignment of sliding organ boundaries we investigated a kernel for discontinuous transformations which is based on *Wendland* functions. Furthermore, we presented a spatially varying kernel for inhomogeneous transformation models. With this kernel, an adapted smoothness level for different anatomical structures such as bones or tissues is achieved. We reached a state-of-the-art registration performance on the 4DCT POPI dataset. By incorporating an inhomogeneous transformation model, we could improve the quality of the registration while keeping the TRE low.

From a theoretical perspective, it would be interesting to further generalize the representer theorem. As in [2] the *standard* representer theorem was proven for nondecreasing  $g$ , the generalized representer theorem may be generalizable to nondecreasing  $g$  as well as nondecreasing  $h$ . For discontinuity preserving problems, direction dependent weighting images for direction dependent regularization [14] would be very interesting. In the experiments, we saw, that there is no winner among our method and pTV, however, we have confirmed a proof of concept. Our current implementation has not been optimized for a particular platform. An efficient implementation which is tailored to GPU architectures would not only speed up the computations but would also enable a systematic parameter search. Finally, a second order stochastic gradient optimizer would potentially increase the convergence speed.

## References

- [1] F. F. Berendsen, A. N. Kotte, M. A. Viergever, and J. P. Pluim. Registration of organs with sliding interfaces and changing topologies. In *SPIE Medical Imaging*, pages 90340E–90340E, 2014. 2
- [2] F. Dinuzzo and B. Schölkopf. The representer theorem for hilbert spaces: a necessary and sufficient condition. In *Advances in neural information processing systems*, pages 189–196, 2012. 2, 7
- [3] M. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE Transactions on Medical Imaging*, 32(7):1239–1248, 2013. 2
- [4] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008. 2
- [5] S. Kiriyanthan, K. Fundana, and P. C. Cattin. Discontinuity preserving registration through motion segmentation with continuous cuts. In *Abdominal Imaging. Computational and Clinical Applications*, volume 7029 of *LNCS*, 2012. 2
- [6] S. Klein, J. P. Pluim, M. Staring, and M. A. Viergever. Adaptive stochastic gradient descent optimisation for image registration. *International journal of computer vision*, 81(3):227–239, 2009. 5
- [7] M. Lüthi, C. Jud, and T. Vetter. Using landmarks as a deformation prior for hybrid image registration. In *Pattern Recognition*, pages 196–205. Springer, 2011. 2

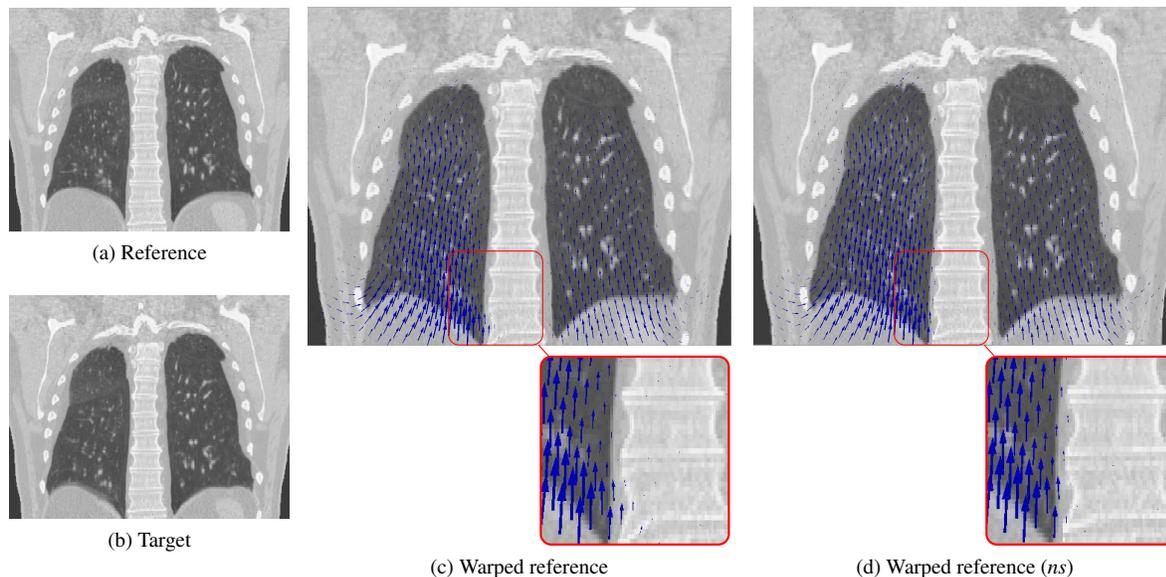


Figure 6: In this figure, we qualitatively compare the registration performance of SKM and *ns*. SKM. In (a)-(b) coronal slice views of the image 6 and 1 are depicted. In (c)-(d), the transformed reference is shown including displacement vectors. The lowest vertebra is less distorted in the nonstationary case, since there, locally smoother transformations are modeled (cf. Figure 5). Thus, the registration quality could be improved without concessions in the TRE (cf. Table 1).

- [8] C. J. Paciorek and M. J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006. 4
- [9] A. Pai, S. Sommer, L. Sorensen, S. Darkner, J. Sporning, and M. Nielsen. Kernel bundle diffeomorphic image registration using stationary velocity fields and wendland basis functions. *Transactions on Medical Imaging*, 2015. 2
- [10] B. W. Papież, M. P. Heinrich, J. Fehrenbach, L. Risser, and J. A. Schnabel. An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration. *Medical image analysis*, 18(8), 2014. 2
- [11] G. K. Rohde, A. Aldroubi, and B. M. Dawant. Adaptive free-form deformation for interpatient medical image registration. In *Medical Imaging 2001*, pages 1578–1587, 2001. 1, 5
- [12] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999. 1, 7
- [13] M. Schmidt, D. Kim, and S. Sra. Projected newton-type methods in machine learning. *Optimization for Machine Learning*, page 305, 2012. 2, 5
- [14] A. Schmidt-Richberg. Sliding motion in image registration. In *Registration Methods for Pulmonary Image Analysis*, pages 65–78. Springer, 2014. 2, 6, 7
- [15] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer, 2001. 3
- [16] B. Schölkopf, F. Steinke, and V. Blanz. Object correspondence as a machine learning problem. In *Proceedings of the 22nd international conference on Machine learning*, pages 776–783. ACM, 2005. 2
- [17] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004. 4
- [18] W. Shi, M. Jantsch, P. Aljabar, L. Pizarro, W. Bai, H. Wang, D. ORegan, X. Zhuang, and D. Rueckert. Temporal sparse free-form deformations. *Medical image analysis*, 17(7):779–789, 2013. 2
- [19] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013. 2
- [20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 2
- [21] J. Vandemeulebroucke, D. Sarrut, and P. Clarysse. The popi-model, a point-validated pixel-based breathing thorax model. In *International conference on the use of computers in radiation therapy*, volume 2, pages 195–199, 2007. 1, 6
- [22] V. Vishnevskiy, T. Gass, G. Székely, and O. Goksel. Total variation regularization of displacements in parametric image registration. In *Abdominal Imaging. Computational and Clinical Applications*, pages 211–220. Springer, 2014. 2, 7
- [23] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4:389–396, 1995. 4
- [24] F. Yan and Y. A. Qi. Sparse gaussian process regression via l1 penalization. In *Proceedings of the 27th International Conference on Machine Learning*, 2010. 2