Unsupervised classification of children’s bodies using currents

Sonia Barahona · Ximo Gual-Arnau · Maria Victoria Ibáñez · Amelia Simó

Received: 24 May 2016 / Revised: 21 February 2017 / Accepted: 2 March 2017 / Published online: 11 March 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract Object classification according to their shape and size is of key importance in many scientific fields. This work focuses on the case where the size and shape of an object is characterized by a current. A current is a mathematical object which has been proved relevant to the modeling of geometrical data, like submanifolds, through integration of vector fields along them. As a consequence of the choice of a vector-valued reproducing kernel Hilbert space (RKHS) as a test space for integrating manifolds, it is possible to consider that shapes are embedded in this Hilbert Space. A vector-valued RKHS is a Hilbert space of vector fields; therefore, it is possible to compute a mean of shapes, or to calculate a distance between two manifolds. This embedding enables us to consider size-and-shape clustering algorithms. These algorithms are applied to a 3D database obtained from an anthropometric survey of the Spanish child population with a potential application to online sales of children’s wear.
1 Introduction

Many problems in medical imaging analysis and computer vision involve the classification of objects based on their shapes or on their sizes and shapes. A significant amount of research and activity has been carried out in recent decades in the general area of shape analysis.

Throughout this work we are going to work with bodies, i.e. geometrical objects with bounded boundaries. Several mathematical frameworks have been proposed in the literature to deal with such objects, three of these being the most widely used. Firstly, functions can be used to represent closed contours of the objects (curves in 2D and surfaces in 3D); secondly geometrical objects can also be treated as subsets of $\mathbb{R}^m$ and, finally, these geometrical objects can be described as sequences of points that are given by certain geometrical or anatomical properties (landmarks). Shapes, in all these settings, are embedded into a space which is not a vector space (in a large number of cases it is a smooth manifold) and on which no natural metric is defined. This makes the definition of statistics particularly difficult; for example, there is no simple explicit way to compute a mean (Pennec 2006).

Recently we have considered the space of planar shapes, represented by simple closed plane curves with a Sobolev-type metric (Gual-Arnau et al. 2015). This space has the property of being isometric to an infinite-dimensional Grassmann manifold of 2-dimensional subspaces, and we have used this isometry to compute geodesics, distances between shapes and mean shapes and we have applied these concepts in order to study different biomedical applications. The corresponding theory for the shape space of surfaces was generalized by Bauer et al. (2011), where shapes are modeled as surfaces in a manifold of surfaces, and distances are computed from the geodesics on this manifold. However, these results consider parameterized curves and surfaces and means are difficult to compute.

In this approach, the contour of each body (curve in $\mathbb{R}^2$, surface in $\mathbb{R}^3$, or hypersurface in $\mathbb{R}^n$), is represented by a mathematical structure named current which allows to identify each shape with a vector field pertaining to a particular Hilbert space. This framework is not limited to a particular kind of data. Indeed, it provides a unifying framework to process any set of points, curves and surfaces or a mixture of these. No hypothesis on the topology of the shapes is assumed. In particular, it is robust to changes of connectivity of the structures. Moreover, it is weakly sensitive to the sampling of shapes and it does not depend on the choice of parameterization. However, the main advantage of this setting is that shapes are embedded into a vector space provided with an inner-product; hence, it is possible to use easy statistical tools; in particular means and distances are easy to compute. Another strength of the method is that shapes do not need a point-to-point correspondence or a map between them; however a weakness of the method is that distances between two shapes are affected
if a rigid motion is applied to one of them; that is the reason because previous registration of shapes is very important. In Ovsjanikov et al. (2012) the authors consider a functional map representation of a bijection mapping between shapes; in Huang et al. (2014) the authors use consistent functional maps to consider a shape difference operator for shapes exploration, which includes shape classification with respect to a region of interest (ROI). In fact, in some applications, such as the use of biometric characteristics to recognize a person, the geometry-feature-based approach analyzes local features, such as nose, eyes..., and their geometric relationships; and, therefore, the human body is characterized by specific and unique geometrical features. However, in our application, we are interested in the global size and shape of a child’s body, then, once fixed the position of the body (arms located at a fixed distance from the body and legs apart a fixed centimeters), the boundary of the child’s body can be considered a surface and we do not need to establish correspondences that are anatomically justified to quantify deformation. Therefore it makes sense to integrate a vector field defined in $R^3$ along the entire surface.

A current is a mathematical object which has been proved relevant for modeling geometrical data like curves and surfaces (Vaillant and Glaunès 2005; Glaunes and Joshi 2006; Durrleman et al. 2009).

From integration on manifolds (Morgan 2008; Lang 1995), if $\Omega_p$ denotes the space of differential $p$-forms in $R^n$, each $p$-dimensional submanifold $X$ in $R^n$ (in particular, $X$ may be the contour of a geometrical object) can be represented by an application that integrates each $p$-form along $X$, i.e. by an application

$$C_X : \Omega_p \rightarrow R$$

$$w \rightarrow \int_X w,$$

such application $C_X$, is called a $p$-current.

In addition, it is possible to associate a subspace of currents to a reproducing kernel Hilbert space (RKHS) by duality. A RKHS is a Hilbert space of mappings which has useful properties. Moreover, these associations allow us to represent each set of piecewise-defined manifolds by a function in a RKHS (Durrleman 2010).

Given a set of geometrical objects (curves or surfaces), our aim is to apply classification techniques developed for Euclidean spaces in order to divide the $m$ objects into $k$ appropriate clusters. A Hilbert space and, even more so, a RKHS can be considered the natural extension of the usual Euclidean spaces $R^n$. The completeness of Hilbert spaces gives a framework in which to work with infinite-dimensional vectors as the limit of finite-dimensional vectors.

This paper arose as the result of an important study conducted by the Valencian Institute of Biomechanics, the ultimate objective of which was to help decision makers (parents/relatives/children) in the size selection process when shopping online for children’s wear.

A 3D anthropometric study of the child population in Spain was carried out for that purpose. After the study was completed, a database was generated consisting of 739 randomly selected Spanish children between 3 and 12 years of age. They were scanned using the Vitus Smart 3D body scanner from human solutions, a non-intrusive laser
system formed by four columns housing the optic system, which moves from head to feet in ten seconds, performing a sweep of the body. Our work focuses on one of the aims of this study: to define of an efficient sizing system.

A standard sizing system classifies a specific population into homogeneous subgroups based on certain key body dimensions (European Committee for Standardization 2002; Chung et al. 2007; Ibáñez et al. 2012). Most of the standard sizing charts propose sizes based on intervals over just two or three anthropometric dimensions. However, correlations between anthropometric measures show great variability in body proportion and as a result it is not possible to cover so many different body morphologies with these kinds of models.

In this paper, instead of using clustering methods to divide the population into sizes by simply using a set of anthropometric variables, we propose to use the body shapes represented by currents and the well-known \textit{k-means} algorithm in the corresponding space.

Shape analysis in the anthropometry context specific to 3D human body has been largely used in recent years. The area of major research is the corresponding to the statistical human modeling (Allen et al. 2003; Pishchulin at al. 2015) whose objective is to create a detailed parametric model of the human body. This model can be used then for a range of applications, such as, ergonomic design or computer animation (Jain et al. 2010; Baek and Lee 2012). The aim of our work is very different, we just look for a good descriptor of shape in order to make a cluster analysis. Our problem is similar to that posed in Vinué et al. (2016), however the approaches are completely different. In that paper the authors represent each shape as a configuration matrix of landmarks.

The original \textit{k-means} algorithm (Steinhaus 1956; Lloyd 1957), endeavors to find a partition such that the sum-of-squares error between the empirical mean of a cluster and the objects in the cluster is minimized. It tries to approximate this optimum \(k\)-partition by iterating. Starting with \(k\) arbitrary initial cluster centers, an initial \(k\)-partition is obtained, assigning each object to its closest cluster center. Next, the new \(k\) cluster centers are re-calculated as the mean of the observations of the clusters resulting from the previous step. This loop is continued until no further changes are made. Many procedures were developed in subsequent decades to improve this classic algorithm; see e.g. Kanungo et al. (2002) and Nazeer and Sebastian (2009). Even though the \textit{k-means} algorithm was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering (Jain 2010).

Our implementations have been written with MATLAB (2014).

The article is organized as follows: Sect. 2 concerns the theoretical concepts of currents and reproducing kernel Hilbert spaces. The practical applications of this theory is explained in Sect. 3. In Sect. 4 the \textit{k-means} algorithm in the RKHS space is introduced. An experimental study with synthetic figures is conducted in Sects. 5 and 6. The application for classifying children’s body shapes is detailed in Sect. 7. Finally, conclusions are discussed in Sect. 8.

Those who are more interested in practical applications that in the mathematical foundation can go directly to Sect. 3, where a representation of plane curves and surfaces as currents is given.
2 From bodies to elements in a reproducing kernel Hilbert space through currents

A detailed study of the theory of currents in this context can be found in Durrlemann (2010). Let $B_1, B_2, \ldots, B_m$ be $m$ bodies in $\mathbb{R}^n$ whose boundaries $S_i = \partial B_i$, ($i = 1, 2, \ldots, m$), are smooth hypersurfaces in $\mathbb{R}^n$. In this section we introduce the theoretical foundations to represent $S_i$ as elements in a reproducing kernel Hilbert space (RKHS). In order to do that, we will first represent the hypersurfaces $S_i$ as geometrical currents.

Currents were introduced by De Rham in 1955 and by the 1960 paper by Federer and Fleming on &ldquo;Normal and Integral Currents&rdquo;, which was awarded the 1986 AMS Steele Prize for a paper of fundamental or lasting importance; but their use in computational anatomy is recent (Glaunès 2005).

Let $\Omega^m_{n-1}$ denote the space of continuous $(n-1)$–forms on $\mathbb{R}^n$. The space of $(n-1)$–currents on $\mathbb{R}^n$ is the topological dual $\Omega^{m}_{n-1}$; i.e. the space of linear and continuous forms on $\Omega^m_{n-1}$; and it is a fact that every hypersurface $S$ of $\mathbb{R}^n$ with a finite volume can be represented by an element of $\Omega^m_{n-1}$. That is, from integration on manifolds (Morgan 2008), we know that any $(n-1)$–dimensional form of $\Omega^m_{n-1}$ can be integrated along the hypersurface $S$, which associates with $S$ a $(n-1)$–current $C_S$ such that:

$$C_S(\omega) = \int_S \omega, \quad \forall \omega \in \Omega^m_{n-1}. \quad (2)$$

The application $S \mapsto C_S$ is injective but not surjective [not all the currents can be represented by integration on a hypersurface (geometrical currents)].

Suppose that the hypersurface $S$ is a parameterized surface $r : D \subset \mathbb{R}^{n-1} \rightarrow \mathbb{R}^n$ with $r(D) = S$, then $r(x) = r(x_1, x_2, \ldots, x_{n-1}) \in S$, and

$$C_S(\omega) = \int_S \omega = \int_D \omega(r(x)) \left( \frac{\partial r}{\partial x_1} \wedge \cdots \wedge \frac{\partial r}{\partial x_{n-1}} \right) dx_1 \cdots dx_{n-1}. \quad (3)$$

This representation is fully geometric in the sense that it only depends on the hypersurface structure and not on the choice of parameterization. Moreover the representation of a surface as a geometrical current distinguishes between isometric hypersurfaces (that is, hypersurfaces obtained as rotations and/or translation). On the other hand, the opposite current-$C_S$ represents the same hypersurface but with the opposite orientation (since the flux through the hypersurface has the opposite sign in this case).

2.1 Vectorial representation of geometrical currents

A form $\omega \in \Omega^m_{n-1}$ can be associated with a vector field $\vec{\omega}$ on $\mathbb{R}^n$ thanks to the isometric mapping between the $(n-1)$–form on $\Omega^m_{n-1}$ and the vectors on $\mathbb{R}^n$. Then,
\[ C_S(\omega) = \int_D \omega(r(x)) \left( \frac{\partial r}{\partial x_1} \wedge \cdots \wedge \frac{\partial r}{\partial x_{n-1}} \right) dx_1 \cdots dx_{n-1} \]
\[ = \int_D \text{Det} \left( \frac{\partial r}{\partial x_1}, \ldots, \frac{\partial r}{\partial x_{n-1}}, \bar{\omega}(r(x)) \right) dx_1 \cdots dx_{n-1}. \]  

(4)

Formally, the association between forms and vectors is given by the Hodge star operator and duality (Do Carmo 2012).

2.2 Particular cases: planar curves and surfaces

As stated in the introduction, we are interested in the particular cases of planar closed curves and compact surfaces (contours of bodies in \( \mathbb{R}^2 \) or \( \mathbb{R}^3 \)).

Let \( \alpha : I = [a, b] \rightarrow \mathbb{R}^2 \) be a parameterized regular oriented simple curve in \( \mathbb{R}^2 \). We associate with \( \alpha \) the function (geometrical current)

\[ C_\alpha(\omega) = \int_a^b \bar{\omega}(\alpha(t)) \cdot \alpha'(t) \, dt, \]  

(5)

where \( \bar{\omega} \) is a vector field in \( \mathbb{R}^2 \) and \( \cdot \) denotes the inner product in \( \mathbb{R}^2 \).

Let \( S \) be an orientable parameterized surface in \( \mathbb{R}^3 \) given by \( r : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3; \) that is, \( r(U) = S \). We associate with \( S \) the function (geometrical current)

\[ C_S(\omega) = \int_U \bar{\omega}(r(u, v)) \cdot (r_u(u, v) \wedge r_v(u, v)) \, du \, dv, \]  

(6)

where \( \bar{\omega} \) is a vector field in \( \mathbb{R}^3 \), \( \cdot \) denotes the inner product in \( \mathbb{R}^3 \), \( r_u = \partial r/\partial u \) and \( r_v = \partial r/\partial v \).

Then, to characterize hypersurfaces (mainly curves and surfaces) from the above expressions, we measure how these integrals vary as the vector field \( \omega \) varies. However, instead of considering all the vector fields, we will define a test space of square-integrable vector fields where \( \omega \) varies. In particular, as in Durrleman (2010), we will choose as the test space a vector-valued reproducing kernel Hilbert space (RKHS).

It is important to note that in this case the application \( S \rightarrow C_S \) will not be injective; that is, the same geometrical current \( C_S \), as a map defined in a RKHS, may represent two different hypersurfaces. Therefore, the choice of the appropriate RKHS will depend on the application proposed.

2.3 Operator-valued kernels and test space vector-valued RKHS

This section gives a definition of a vector-valued reproducing kernel Hilbert space from the Riesz representation theorem and studies its properties.

The abstract theory of scalar-valued RKHS was developed by Aronszajn (1950). A scalar-valued RKHS is a Hilbert space of functions \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) with some practical properties. In recent years, the study of RKHS has been extended to vector-valued
Unsupervised classification of children’s... 371

functions (the space contains vector fields from \( \mathbb{R}^n \) to \( \mathbb{R}^n \)) (see Carmeli et al. 2006; Micchelli and Pontil 2005; Caponnetto et al. 2008) and it has now become a widely studied theory.

Let \( \mathcal{L}(\mathbb{R}^n) \) be the Banach space of bounded linear functions from \( \mathbb{R}^n \) to \( \mathbb{R}^n \).

**Definition 2.1** Let \( W \) be a Hilbert space of vector fields from \( \mathbb{R}^n \) to \( \mathbb{R}^n \). An operator \( K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathcal{L}(\mathbb{R}^n) \) is said to be an operator-valued reproducing kernel (rk) associated with \( W \) if

1. for every \( x, \alpha \in \mathbb{R}^n \), \( K(\cdot, x)(\alpha) \in W \) (where \( K(\cdot, x)(\alpha)(y) = K(x, y)(\alpha) \in \mathbb{R}^n \), \( \forall y \in \mathbb{R}^n \)) and,
2. \( K \) satisfies the “reproducing property”; that is, \( \forall \omega \in W \) and \( x, \alpha \in \mathbb{R}^n \)

\[
\omega(x) \cdot \alpha = \langle K(\cdot, x)(\alpha), \omega \rangle_W
\]

**Definition 2.2** Let \( W \) be a Hilbert space of vector fields from \( \mathbb{R}^n \) to \( \mathbb{R}^n \). \( W \) is a vector-valued RKHS if there is an operator-valued reproducing kernel (rk) associated with \( W \).

The next theorem (see below) is a sort of converse to this: if a function \( K \) is both symmetric and positive definite, then there is a Hilbert space of vector fields from \( \mathbb{R}^n \) to \( \mathbb{R}^n \) for which it is a reproducing kernel.

**Definition 2.3** A function \( K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathcal{L}(\mathbb{R}^n) \) is said to be an operator-valued positive definite and self-adjoint kernel if for each pair \( (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \), \( K(x, y) \in \mathcal{L}(\mathbb{R}^n) \) is a self-adjoint operator and

\[
\sum_{i,j=1}^{N} \alpha_i \cdot K(x_i, x_j)(\alpha_j) \geq 0,
\]

for every finite set of points \( \{x_i\}_{i=1}^{N} \in \mathbb{R}^n \) and \( \{\alpha_i\}_{i=1}^{N} \in \mathbb{R}^n \).

**Theorem 2.4** If \( K \) is an operator-valued positive definite and self-adjoint kernel, then there is a unique RKHS, \( W \), such that \( K \) is the operator-valued reproducing kernel (rk) associated with \( W \).

It is essential to bear in mind that the proof of this theorem is based on constructing the space \( W \) through the completion of \( H_0 := \text{span}\{K(\cdot, x)(\alpha)/x, \alpha \in \mathbb{R}^n\} \subset W \).

For \( \omega_1 = \sum_{i=1}^{N_1} K(\cdot, x_i)(\alpha_i), \omega_2 = \sum_{j=1}^{N_2} K(\cdot, x_j)(\beta_j) \in H_0 \), define

\[
\langle \omega_1, \omega_2 \rangle_{H_0} := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \alpha_i \cdot K(x_i, y_j)(\beta_j). \quad (7)
\]

The vector-valued RKHS \( W \) associated with the kernel \( K \), that, from now on, will be denoted as \( H_K \), is the closure of \( H_0 \), that is, the span of vector fields of the form \( K(x, \cdot)(\alpha) \) for every \( x \in \mathbb{R}^n \) and \( \alpha \in \mathbb{R}^n \) is dense in \( H_K \). For this reason, the inner
product in $H_K$, between $\omega_1, \omega_2 \in \overline{H_0}$ is the limit of expression (7) when $N_1, N_2$ tend to infinity.

Having established the vector-valued RKHS test space $H_K$, currents will be evaluated in $H_K$; that is, the space of currents considered is $H_K^*$ (it contains the continuous linear functions from $H_K$ to $\mathbb{R}$) which includes the geometrical currents as a subset. The space of currents $H_K^*$ is a vector space with the operations sum ($+$) and product ($\cdot$) as in a standard space of functions (Durrleman 2010).

Therefore, the idea is to build the vector space spanned by the vector fields of the form $K(x, \cdot)(\alpha)$ and to make this space complete by adding the limit of every Cauchy sequence to it. This construction make it possible to process discrete meshes of surfaces and continuous surfaces (limit of such a finite combination) in the same setting.

### 2.4 Curves and surfaces as elements in a vector-valued RKHS

We are now going to use the properties of the RKHS, $H_K$ in order to rewrite the geometrical currents associated with curves in $\mathbb{R}^2$ and surfaces in $\mathbb{R}^3$.

Let $\omega$ be a vector field in $H_K(\mathbb{R}^2, \mathbb{R}^2)$. Then, by using Eq. (5), the geometrical current associated with a curve $\alpha$ becomes:

$$C^K_\alpha(\omega) = \int_a^b \tilde{\omega}(\alpha(t)) \cdot \alpha'(t) \, dt = \langle \int_a^b K(\alpha(t), \cdot)(\alpha'(t)) \, dt, \omega \rangle_{H_K},$$

and by Eq. (6), the geometrical current associated with the parameterized surface $S = r(U)$ is:

$$C^K_S(\omega) = \int_U \tilde{\omega}(r(x)) \cdot \eta(x) \, dx = \langle \int_U K(r(x), \cdot)(\eta(x)) \, dx, \omega \rangle_{H_K}$$

where $x = (u, v), \, dx = du \, dv, \, \eta(x) = r_u \wedge r_v$ is the orthogonal vector to the surface $S$ at the point $r(x)$, and $\omega$ and $\int_S K(r(x), \cdot)(\eta(x)) \, dx \in H_K^*(\mathbb{R}^3, \mathbb{R}^3)$.

Until now, each hypersurface has been associated with an element in the vector space $H_K^*$. However, the Riesz Fréchet Theorem (Conway 2013) establishes that there is an isometric, linear, bijective mapping $\mathcal{L}_{H_K} : H_K \rightarrow H_K^*$, defined by $\mathcal{L}_{H_K}(\omega)(\omega') := \langle \omega, \omega' \rangle_{H_K}, \forall \omega, \omega' \in H_K$. As a consequence, the space $H_K^*$ is isometric to $H_K$, and then each hypersurface can be associated with a vector field in $H_K$.

Therefore, as a result of the Riesz-Fréchet Theorem, it is possible to represent the parameterized curve $\alpha$ defined in $[a, b]$, by an element in $H_K(\mathbb{R}^2, \mathbb{R}^2)$; that is

$$\alpha \longrightarrow C^K_\alpha(\omega) = \langle \int_a^b K(\alpha(t), \cdot)(\alpha'(t)) \, dt, \omega \rangle_{H_K} \cong \int_a^b K(\alpha(t), \cdot)(\alpha'(t)) \, dt,$$

(10)
where \( \cong \) denotes the isometric element from the Riesz-Frèchet Theorem, and to represent a parameterized surface \( S \) by

\[
S \rightarrow C^K_S(\omega) = \int_U \bar{\omega}(r(x)) \cdot \eta(x) \, dx \\
= (\int_U K(r(x), \cdot)(\eta(x)) \, dx, \omega)_{H_K} \cong \int_U K(r(x), \cdot)(\eta(x)) \, dx
\]

(11)

Consequently, we deal with curves and surfaces as vector fields in \( H_K \). Thus, the distance between two surfaces (or curves) is defined as the distance between the corresponding elements in \( H_k \); that is, if \( \phi_1 \) and \( \phi_2 \) are two elements in the RKHS associated, for instance, with two surfaces \( S_1 \) and \( S_2 \); then:

\[
d(S_1, S_2)^2 = d(\phi_1, \phi_2)^2 = (\phi_1 - \phi_2, \phi_1 - \phi_2)_{H_k} \\
= (\phi_1, \phi_1)_{H_k} - 2(\phi_1, \phi_2)_{H_K} + (\phi_2, \phi_2)_{H_K}. \quad (12)
\]

2.5 Choice of the operator-valued reproducing kernel

The choice of the kernel determines the vector-valued RKHS and especially its metric. The choice of this metric is therefore crucial and must be adapted to every particular application.

Based on this, we are going to use a particular class of operator-valued kernel \( K \) defined as follow:

**Definition 2.5** Let \( D \subseteq \mathbb{R}^n \) be a non-empty subset, and \( K : D \times D \rightarrow \mathcal{L}(\mathbb{R}^n) \) an application. Then, for each \( x, y \in D \), the operator \( K(x, y) \) is defined by

\[
K(x, y) : \mathbb{R}^n \rightarrow \mathbb{R}^n \\
\alpha \rightarrow K(x, y)(\alpha) := k(x, y) \cdot \alpha
\]

(13)

where \( k : D \times D \rightarrow \mathbb{R} \) is a symmetric and positive semi-definite function, (i.e. \( k(x, y) = k(y, x) \forall x, y \in D \) and \( \sum_{i,j} a_i k(x_i, x_j)(a_j) \geq 0 \) for finite sets \( \{a_i\} \in \mathbb{R}, \{x_i\} \in D \)).

**Proposition 2.6** The operator-valued \( K \) established in the previous definition is well defined, symmetric and positive semi-definite, so there is a unique RKHS vector-valued \( H_K(D, \mathbb{R}^n) \subseteq W := \{\omega : D \rightarrow \mathbb{R}^n\} \) (or simply \( H_K \)) with \( K \) as its operator-valued reproducing kernel.

**Proof** Given \( (x, y) \in D \times D \), the operator \( K(x, y) \) is obviously linear. In addition, \( K(x, y) \) is bounded because its norm is bounded by \( |k(x, y)| \):

\[
\|K(x, y)\| = \sup\{\|K(x, y)(\alpha)\|_{\mathbb{R}^n} : \alpha \in \mathbb{R}^n, \|\alpha\|_{\mathbb{R}^n} \leq 1\} \leq |k(x, y)|.
\]

Moreover, as \( k \) is symmetric and positive semi-definite, \( K \) immediately has these properties. Hence, \( K \) is an operator-valued reproducing kernel and there is a unique RKHS vector-valued \( H_K(D, \mathbb{R}^n) \) with \( K \) as its rk. \( \Box \)
Although it is not known how to choose the “best” kernel for a given application (see Appendix B of Durrleman 2010), translation-invariant isotropic kernels, defined from translation-invariant isotropic scalar kernels of the form \( k(x, y) = k(\|x - y\|_{\mathbb{R}^n}) \) are often used. In these cases, if we apply the same rigid transformation \( \rho \) to two surfaces \( S_1 \) and \( S_2 \) (or curves), we have that \( d(S_1, S_2) = d(\rho(S_1), \rho(S_2)) \). However, the proposed measure would be affected if we apply a rigid transformation only to one of both surfaces.

In particular the Gaussian function (also called the Gaussian kernel)

\[
\begin{align*}
  k(x, y) := e^{-\frac{\|x - y\|_{\mathbb{R}^n}^2}{\lambda^2}}.
\end{align*}
\]

where \( \lambda > 0 \) is a scale parameter (bandwidth), defines an operator-valued \( K : D \times D \to \mathcal{L}(\mathbb{R}^n) \) that is particularly important in the literature and it is called Vector-Valued Gaussian. The vector-valued RKHS \( H_K(D, \mathbb{R}^n) \) defined by it has the following expression (Quang et al. 2010)

\[
\begin{align*}
  H_K(D, \mathbb{R}^n) = \left\{ f \in C_0(D, \mathbb{R}^n) \cap L^2(D, \mathbb{R}^n) : \int_{\mathbb{R}^n} e^{-\frac{\lambda^2 \|\xi\|^2_{\mathbb{R}^n}}{4}} \|\hat{f}(\xi)\|^2_{\mathbb{R}^n} d\xi < \infty \right\}
\end{align*}
\]

where \( \hat{f} \) is the Fourier transform of \( f \).

**Proposition 2.7** Given \( D \subseteq \mathbb{R}^n \) a non-empty subset; \( K_\lambda : D \times D \to \mathcal{L}(\mathbb{R}^n) \) the operator defined as in Eq. (13) through the Gaussian kernel \( k \) with parameter \( \lambda \) (Eq. 14), and \( H_{K_\lambda} \) the RKHS established by \( K_\lambda \), if \( \lambda_1, \lambda_2 \in \mathbb{R} : 0 < \lambda_1 \leq \lambda_2 \), then \( H_{K_{\lambda_2}} \subseteq H_{K_{\lambda_1}} \).

**Proof** If \( f \) was in \( H_{K_{\lambda_2}} \), using the space expression (15), the integral with \( \lambda_2 \) would be finite and larger than the integral with parameter \( \lambda_1 \),

\[
\int_{\mathbb{R}^n} e^{-\frac{\lambda_1^2 \|\xi\|^2_{\mathbb{R}^n}}{4}} \|\hat{f}(\xi)\|^2_{\mathbb{R}^n} d\xi \leq \int_{\mathbb{R}^n} e^{-\frac{\lambda_2^2 \|\xi\|^2_{\mathbb{R}^n}}{4}} \|\hat{f}(\xi)\|^2_{\mathbb{R}^n} d\xi < \infty
\]

Hence, \( f \in H_{K_{\lambda_1}} \). In conclusion, the smaller the value of \( \lambda > 0 \), the greater the space \( H_{K_\lambda} \) established.

The Gaussian kernel will be used in our application of Sect. 7. In Sect. 5, it will be compared experimentally to another popular translation-invariant isotropic kernel.

**Remark 2.8** Choice of parameter \( \lambda \)

Since, the smaller the value of \( \lambda > 0 \), the greater the space \( H_{K_\lambda} \) established, (proposition 2.7). It is important to note that the larger \( H_{K_\lambda} \), the better the differentiation between geometrical data, because since the test space would be greater, it would be more likely to identify different values that the current has in vector fields. Thus, the
smaller the value of $\lambda > 0$, the greater the precision for characterizing geometrical data. For this reason, if $\lambda$ is too small, the distance in RKHS detects tiny geometrical details and too much noise could be captured. In conclusion, it is essential to choose a suitable parameter balancing the two previous ideas, so $\lambda > 0$ should be the typical scale at which the vector fields $\omega \in H_K(D, \mathbb{R}^n)$ may vary spatially.

3 Practical application

In practical applications curves and surfaces are just known at a finite number $p$ of points. In this section we see how to use the theory of the previous section in this situation.

Consider now a curve $\alpha$ known at a finite number, $p$, of points $\{t_1 < t_2 < \cdots < t_p\}$, that constitute a partition of the interval $[a, b]$. Let $y_j = \alpha(t_j) \in \mathbb{R}^2 \quad \forall j = 1, \ldots, p$, let $x_j \in \mathbb{R}^2$ denote the center of the segment $[y_j, y_{j+1}]$ and let $\tau_j = y_{j+1} - y_j$ be an approximation of the tangent vector (the finer the partition the better). Then, using the results of the previous section we can represent it by:

$$\alpha \longrightarrow C^K_\alpha \cong \int_a^b K(\alpha(t), \cdot)(\alpha'(t))dt = \int_a^b k(\alpha(t), \cdot) \cdot \alpha'(t)dt = \sum_{j=1}^p k(x_j, \cdot) \cdot \tau_j,$$

(16)

where $k$ is a scalar kernel function as for example the one given in Eq. 14.

So, in practical applications each curve $\alpha$ will be represented by a finite addition which is an approximation to the vector field (Fig. 1).

If we suppose now that we have a triangulation of a surface $S$ where each triangle is $y_jy_{j+1}y_{j+2}$, $x_j = \frac{1}{3}(y_j + y_{j+1} + y_{j+2})$ and $\tau_j = \frac{1}{2}(y_{j+1} - y_j) \wedge (y_{j+2} - y_j)$ ($\tau_j$ is the normal vector to the triangle, whose norm encodes the area of the triangle); then, $S$ is represented by the vector field:

![Fig. 1 Curve in $\mathbb{R}^2$ and its elements](image)
\[ S \rightarrow C_{S}^{K}(\omega) \cong \int_{U} K(r(x), \cdot)(\eta(x)) \, dx = \int_{U} k(r(x), \cdot) \cdot \eta(x) \, dx \approx \sum_{j=1}^{p} k(x_{j}, \cdot) \cdot \tau_{j}. \] (17)

The finite addition tends towards the integral as the mesh is refined. This finite addition is an approximation to the vector field and it will be the representation which will be used as a consequence of its computational simplicity.

Using this finite approximation to \( \varphi_{1} = \sum_{j=1}^{p_{1}} k(x_{j}^{1}, \cdot) \cdot \tau_{j}^{1} \) and \( \varphi_{2} = \sum_{j=1}^{p_{2}} k(x_{j}^{2}, \cdot) \cdot \tau_{j}^{2} \) we have:

\[ \langle \varphi_{1}, \varphi_{2} \rangle_{H_{K}} = \sum_{i=1}^{p_{1}} \sum_{j=1}^{p_{2}} \tau_{i}^{1} \cdot k \left( x_{i}^{1}, x_{j}^{2} \right) \cdot \tau_{j}^{2}. \] (18)

And, as a result (Eq. 12):

\[ d(\varphi_{1}, \varphi_{1})^{2} = \sum_{i=1}^{p_{1}} \sum_{j=1}^{p_{1}} \tau_{i}^{1} \cdot k \left( x_{i}^{1}, x_{j}^{1} \right) \cdot \tau_{j}^{1} \]

\[ + \sum_{i=1}^{p_{2}} \sum_{j=1}^{p_{2}} \tau_{i}^{2} \cdot k \left( x_{i}^{2}, x_{j}^{2} \right) \cdot \tau_{j}^{2} - 2 \sum_{i=1}^{p_{1}} \sum_{j=1}^{p_{2}} \tau_{i}^{1} \cdot k \left( x_{i}^{1}, x_{j}^{2} \right) \cdot \tau_{j}^{2}. \] (19)

This distance between currents does not require any correspondence among points of shapes and it is unnecessary to consider the same number of points in each shape to compute the distance. However, hereinafter and in our applications, we use a fixed number of points \( p \) for all shapes in the sample. These points are obtained from a triangularization of the surfaces (Fig. 2).

In the Sect. 6 we perform an experimental study to analyze the influence of the density of points on the surface in the distance.

Our method assumes that all the children’s bodies are registered. Although the problem of shape registration based on currents may be formulated using flows of diffeomorphisms (see Durrleman 2010), given the characteristics of the children’s bodies and the fact that the all children were scanned in the same position, we use a simple criteria for registration. First, children were scanned in the Valencian Institute of Biomechanics with the Vitus XXL scanner in the same two standing postures. 3D scan data from all bodies were processed for the creation of posture-harmonised homologous models and were aligned in the \( Z \) axis. To finish the registration, a point in the \( XY \)-plane is searched to center all the models. Since the models are standing with feet on the floor, separated to shoulder width, we have chosen this point as the Euclidean mean of the coordinates of the points which are on the outline of the ankles.

Moreover, the sample mean in the vector-valued RKHS \( H_{K} \) is calculated in the same way as in Euclidean spaces (Hsing and Eubank 2015). Given a sample \( \varphi_{1}, \ldots, \varphi_{m} \) the sample mean is:
Fig. 2 Triangulated surface in \( \mathbb{R}^3 \) and its elements

\[
\bar{\varphi} = \frac{\sum_{i=1}^{m} \varphi_i}{m} = \frac{\left( \sum_{i=1}^{m} \sum_{j=1}^{p} k \left( x_j^i, \cdot \right) \tau_j^l \right)}{m}, \tag{20}
\]

\( \bar{\varphi} \in H_k \); however, in general, \( \bar{\varphi} \) is not a geometrical current associated with a surface.

The distance between surfaces, obtained from the distance between the corresponding geometrical currents, gives a global estimate of the shape dissimilarity between objects. This distance will be used for the definition of an efficient sizing system for the child population, in which a global dissimilarity shape measure between bodies that does not highlight where the differences occur locally is needed.

4 \( k \)-means algorithm in the RKHS space

In this section we review the classic \( k \)-means partitioning algorithm and we comment how adapt it to the RKHS space introduced in the previous section.

Given \( m \) data points \( x_1, \ldots, x_m \) in \( \mathbb{R}^d \) and a \( k \)-partition \( \mathcal{C} = (C_1, \ldots, C_k) \) of the set \( \mathcal{O} = \{1, \ldots, m\} \) of underlying objects, with non-empty classes, let:

\[
W(\mathcal{C}) = \sum_{i=1}^{k} \sum_{l \in C_i} \| x_l - \bar{x}_i \|^2, \tag{21}
\]

where \( \bar{x}_i \) denotes the centroid of the data points.

As is well known, the classic \( k \)-means clustering approach looks for a \( k \)-partition \( \mathcal{C} \) of \( \mathcal{O} \) such that a minimum value of \( W(\mathcal{C}) \) is reached. This one-parameter optimization problem is equivalent to the two-parameter optimization problem:
where minimization is also w.r.t. all vectors \( Z = (z_1, \ldots, z_k) \) of \( k \) points \( z_1, \ldots, z_k \) from \( \mathbb{R}^d \) (class representatives, class prototypes).

The \( k \)-means algorithm tries to approximate an optimum \( k \)-partition by iterating the partial minimization steps (Bock 2007):

**Algorithm 1**

1. Given an initial partition \( C(0) \), obtain the centroid vector \( Z(0) = (\bar{x}_1^0, \ldots, \bar{x}_k^0) \). Set \( i = 1 \).
2. Given a centroid vector \( Z(i-1) \), obtain \( C(i) \) minimizing Eq. (22) with respect to \( C \), assigning each point to the class whose centroid has the minimum Euclidean distance to it;
3. Given \( C(i) \), minimize Eq. (22) with respect to \( Z \), obtaining the new centroid vector \( Z(i) = (\bar{x}_1^i, \ldots, \bar{x}_k^i) \), the sample means.
4. Set \( i = i + 1 \) and go to STEP 2 until convergence is reached.

By construction, this algorithm yields a sequence \( Z(0), C(1), Z(1), C(2), Z(2), \ldots \) of centroids and partitions with decreasing values of the objective function (Eq. 22) that typically converges towards a (typically local) minimum value.

In this paper, our sample is a set of vector fields \( \varphi_j = \sum_{i=1}^p K(x_j^i, \cdot)(\tau_j^i) \in H_K \) \((j = 1, \ldots, m)\) which represent geometrical data by currents. The adaptation of the above algorithm to the RKHS space of these vector fields is straightforward.

In this space the proximity between elements in the sample (Euclidean distance in Eqs. 21, 22) is measured through distance of vector fields in \( H_K \). Thus, in the functions to minimize in the previous Algorithm, the Euclidean distance must be replaced by the distance given in Eq. 19.

The sample means of STEP 3 in the previous algorithm are obtained using Eq. 20.

### 5 Experimental 2D study

In this section we study the performance of our procedure in a shape clustering problem using a known database of synthetic figures called the MPEG7CEShape-1 PartB database. It includes binary images grouped into categories like cars, faces, watches, horses and birds with images corresponding to the same item, but showing noticeably different shapes.

To perform this experimental study, three classes from this database of synthetic figures were considered: cars, faces and watches. Each class contains 20 elements (binary images) except the watch class where two of them were rejected (watch-2 is an atypical element because of its very large size and watch-8 is turned and our theoretical framework considers size and shape). The figures were centered and the contour \( \alpha_k \) from each of them defined an oriented smooth curve which was discretized by 100 points \( \{y_j^k\}_{j=1}^{100} \) \((y_1^k = y_{100}^k) \) for \( k = 1, \ldots, 58 \). Moreover, face figures were
rotated by $90^\circ$, in order to keep a common horizontal orientation in all synthetic figures (establishing the correspondence with children database, where all the elements are registered and have the same position). For each $k \in \{1, \ldots, 58\}$, from \{$y^k_j\}_{j=1}^{100}$, we defined $x^k_j = \frac{1}{2}(y^k_j + y^k_{j+1})$ the centers of the segments and the vectors $\tau^k_j = y^k_{j+1} - y^k_j$, $\forall j = 1, \ldots, 99$, which define the vector field $\sum_{j=1}^{99} K(x^k_j, \cdot)(\tau^k_j)$ in $H_K(D, \mathbb{R}^2)$.

In this experimental study, we were interested in analyzing similar situations to the ones that will appear in our real application in Sect. 7. We therefore considered two scenarios. In the first, all the 58 synthetic figures were contracted or expanded to reach the same length in the $X$ axis, establishing the similarity between this length and the height of a child. Figure 3 shows an example of an object from each class in this first situation. The points \{$x^k_j\}_{j=1}^{99}$ are plotted in black and the vectors \{$\tau^k_j\}_{j=1}^{99}$ from each curve are plotted in different colors.

In the second scenario, half of the synthetic figures from each category were enlarged by a scale factor of 1.5. In this case there were two different “heights” for each class of figures. Moreover, each figure of the sample was multiplied by a random coefficient ranging between 1 and 1.1, in order to change the “height” of the figures somewhat.

Figure 4 shows an example of an object from each group in the second situation, in which there are two “heights” from each class of shapes.

The $k$-means algorithm presented in Sect. 4 was then applied to both scenarios, choosing as the value of the parameter $\lambda$ the standard deviation of the points \{$x^k_j\}_{j=1}^{99}$, $k = 1, \ldots, 58$ which define the curves of the sample of each scenario. So we defined the Gaussian kernels with $\lambda = 50.16$ in the first scenario and with $\lambda = 67.35$ in the second. In the first scenario the $k$-means algorithm recovered the three groups in...
accordance with the categories of figures in the database. In the second scenario, the k-means algorithm with \( k=6 \), grouped together the figures of the database with the same “height” and shape, as hoped.

In order to check the adequacy of the distance proposed, a multidimensional scaling analysis (MDS) (Cox and Cox 2000) has been carried out. The MDS algorithm is a data reduction descriptive procedure to display the information contained in a \((m \times m)\)-distance matrix, \( D \), in a low-dimensional space such that the between-object distances are preserved as well as possible. Then, the method looks for a set of orthogonal variables \( \{y_1, \ldots, y_p\}, \ p < m \) such that the Euclidean distance of the elements regarding these variable are as close as possible to the distances given in the original matrix \( D \). In Fig. 5, MDS has been applied on the distance matrices computed in both scenarios. Clearly distinguished groups can be observed in both graphs.

### 5.1 Sensitivity/robustness of the metric

To evaluate the influence of the parameter \( \lambda \) on the metric and on the k-means results, an experimental study has been carried out with the images of this experimental 2D study. Different \( \lambda \)-values are considered in the Gaussian kernel to compute the dissimilarities between the different shapes. Table 1 shows the errors obtained in the classification for different values of \( \lambda \) and for each scenario. Because this is a non-supervised classification problem, in order to measure the goodness of the classification we consider the minimum classification error when considering each class as true class.

The experimental results verify the idea described in Remark (2.8): the suitable parameter \( \lambda \) is chosen from a balance between small and large values. Groups in
Unsupervised classification of children’s...

Fig. 5  MDS applied on the matrix distance computed in the first scenario, and in the second scenario. 

(a) λ = 52.9492, (b) λ = 67.5789

Table 1  k-means results in the experimental 2D study for different values of λ using the Gaussian kernel

| Scenario 1 | Scenario 2 |
|------------|------------|
| Error (%)  | Error (%)  |
| λ = 1  | 48.27     | λ = 2  | 48.27     |
| λ = 2  | 1.72      | λ = 5  | 6.90      |
| λ = 5  | 0         | λ = 10 | 6.90      |
| λ = 25 | 0         | λ = 20 | 0         |
| λ = 52.9492 | 0     | λ = 50 | 0         |
| λ = 75 | 0         | λ = 67.5789 | 0 |
| λ = 100 | 0        | λ = 75 | 0         |
| λ = 500 | 0        | λ = 80 | 6.90      |
| λ = 10^3 | 0        | λ = 100 | 6.90      |
| λ = 10^6 | 0        | λ = 10^3 | 8.62 |

Scenario 1 are clearly distinguished and only very small values of λ give a positive error. However, Scenario 2 is more complex and a smaller interval of λ-values than in the first scenario gives 0% error. It is important to note that too small values of λ tend to make orthogonal parts of the contours which are far apart; that is, if ϕ_1 is the vector field associated to a piece of the first curve, and ϕ_2 is the vector field associated to a piece of the second curve, which are at a great distance compared to λ, then ||ϕ_1 − ϕ_2||_{HK}^2 ∼ ||ϕ_1||_{HK}^2 + ||ϕ_2||_{HK}^2, and the metric does not capture any shape differences ( < ϕ_1, ϕ_2 >_{HK} ∼ 0).

Figure 6 shows the results of the MDS, applied on the matrix distance computed in the first scenario, where 20 faces, 20 cars and 18 clocks are considered. It can be seen that the distance is appropriate for performing a cluster analysis in all the analyzed cases, for λ ≥ 5. Figure 7 shows the results of the MDS, applied on the matrix distance computed in the second scenario.
Fig. 6 MDS applied on the matrix distance computed in the first scenario, with different $\lambda$-values. a $\lambda = 2$, b $\lambda = 5$, c $\lambda = 25$, d $\lambda = 100$, e $\lambda = 500$, f $\lambda = 1,000,000 $
Finally, we repeat the experimental 2D study in both scenarios by using the Epanechnikov kernel

\[ k_E(x, y) = \left( 1 - \frac{\|x - y\|^2}{\lambda^2} \right) 1_{\{\|x - y\| \leq \lambda\}}, \]  

(23)

with different values of \( \lambda \) (\( 1_{\{\ldots\} \) is the indicator function). This is also a translation-invariant isotropic kernel. Table 2 shows the errors obtained in the classifications by
using the Epanechnikov kernel. Figures 8 and 9 show the results of the MDS, applied on the matrix distance computed using this kernel in both scenarios for several $\lambda$-values.

By using this kernel, there exists again an interval of values of $\lambda$ where the k-means algorithm recovers the groups in accordance with the categories of figures in the database.

### 6 Experimental 3D study

In this section we describe the experiments carried out with a database of 3D figures, to test the same situations as in the previous 2D database. Now three classes of 3D objects: ellipsoids, spheres and pears are considered. Each class contains 10 elements centered in the origin and the contour from each of them is defined by an oriented smooth triangulated surface immersed in $\mathbb{R}^3$. Each surface $k$, is defined from 10,000 triangles with barycenter $x^k_j$ and normal vectors $\tau^k_j$, $j = 1, \ldots, 10,000$, being $\sum_{j=1}^{10,000} K(x^k_j, \cdot)(\tau^k_j)$ the vector field associated with each surface.

We consider the same scenarios as above. In the first, (Fig. 10) all figures have approximately the same length in the Y axis (same “height”).

Figure 11 shows the second scenario, in which there are two “heights” for each class of 3D objects.

The k-means algorithm was applied in both scenarios. Once again the value of parameter $\lambda$ was chosen as the standard deviation of the points $\{x^k_j\}_{j=1}^{10,000}$, $k = 1, \ldots, 30$ (following the criterion also used for the 2D database). The values obtained were $\lambda = 0.6237$ in the first scenario and $\lambda = 0.7934$ in the second. Once again, the k-means algorithm recovered the three groups with the three different geometrical objects in the first scenario, and correctly segmented the figures of the second scenario into six groups, in relation to their “height” and shape.
6.1 Robustness regarding the density of points on the surface

In this subsection we study the influence of the number of triangles which define each surface on the distance between them. A pear of the experimental 3D study is triangularized in four different ways. Each one of the triangulations $T_0$, $T_1$, $T_2$ and $T_3$ divides the pear in a different number of triangles: 10,000, 5600, 3600 and 3000, respectively. In Fig. 12 two of the triangulations are shown.

By using Eq. (19), we compute the distances between the pear with the original triangulation $T_0$, (10,000 triangles), and the other three generated triangulations, $T_1$, $T_2$ and $T_3$. The results appear in Table 3:

Taking into account that the minimum distance between two different pears in the 3D experimental study is 0.2; the distances obtained in Table 3 can be considered irrelevant.
Fig. 9 MDS applied on the matrix distance computed in the second scenario, with the Epanechnikov kernel and different $\lambda$-values. a $\lambda = 10$, b $\lambda = 50$, c $\lambda = 100$, d $\lambda = 1000$

Fig. 10 An object from each class from the 3D synthetic database (Scenario 1)

7 Application to clustering children’s body shapes

The aim of this section is to show how the aforementioned methods can be used to develop a more efficient apparel sizing system that can increase accommodation of the population, taking into account the child’s body shape. At this point, some comments
have to be made about the metric used on currents. First, we have to note that it gives a global measure of shape dissimilarity, though, it cannot distinguish between local shape differences occurring at different places. This is a drawback shared with many other common and known distances, such as the Euclidean distance or the $L^2$ metric in the $L^2$ space of functions. To solve this drawback, Durrleman (2010) and Durrleman et al. (2009) combine it with deformation models to better localize shape differences. Nevertheless, this limitation does not affect this application. We have to take into account that when allocating the size of a garment for a child, it is more important the overall shape of the child’s body that particular local shape features occurring at different places. For example, a child with a small measure of chest and a large waist measurement usually wears the same size as one child with a big measure of chest and a small waist measurement. Furthermore, in the case of adults, the bodies usually present more localized differences in specific areas (breast, stomach, …) that in the case of children. For these reasons, this global measure works properly in our application. In fact, the results of our clustering algorithm could then be used, in a future application, to choose the most suitable size in a potential online sales application.

Before presenting the application, it is fundamental to emphasize a consideration about the current children sizing system, which is based on the sex and the height or the age of buyers, i.e. when a child wants to buy a T-shirt, he/she has to purchase the
size associated with his/her height. It is important to observe that this size is designed for a specific body shape; however, there is a great deal of variability in body shapes of children with the same height. For example, if the size associated with the child’s height does not fit as a consequence of his/her body shape, he/she has to buy the previous or next size of T-shirt, which will probably be too short/long for him/her. In conclusion, it is essential to create a new sizing system which takes into account both: size and shape of the body.

A randomly selected sample of Spanish children aged 3–10 years old were scanned using a Vitus Smart 3D body scanner from human solutions. Each child was situated standing and looking forward; and the body shape was stored as a set of 3075 3D-landmarks (homologous points on his/her surface). The children were asked to wear a standard white garment in order to standardize the measurements. From the 3D mesh, several anthropometric measurements were calculated semi-automatically by combining automatic measurements based on geometric characteristic points with a manual review.

For each child, the landmarks observed made it possible to define an oriented smooth triangulated surface immersed in $\mathbb{R}^3$, with a total of 4668 triangles and the $k$-means algorithm adapted to the RKHS space can be applied.

The question of the number $k$ of clusters to choose is a difficult problem in data clustering and in particular, in our application. From the point of view of defining a sizing system, it is not profitable to design many sizes because it would be very expensive for apparel companies. On the other hand, because the objective of an apparel sizing system is to accommodate as large a percentage of the population as possible, it is not reasonable to define too few sizes either.

As there is a different size system for each sex, in order to illustrate our procedure the subset of the girls older than 6 was chosen from the whole data set. Children younger than 6 have difficulties in maintaining a standard position during the scanning process, so they were excluded from our data set. This selection results in a sample of size 195. According to the European standard UNE-EN 13402-3, this age range has 4 different sizes associated with it (1190–1250, 1250–1310, 1310–1370, 1370–1430 mm).

Consequently, we propose two possible sizing system. The first, aims to divide each height range suggested by the UNE-EN 13402-3 standard into two groups (two sizes). We will thus obtain two different sizes for girls with different body shapes within a common height range of. By using this model, most clothes should fit their buyers. The second sizing system, divides the sample into a lower number of groups (a similar number to that proposed by UNE-EN 13402-3), but bearing in mind the shape and size of the body, not just the height of the children.

Therefore, the body contour from each girl in our data-set is represented by an oriented triangulated smooth surface. For the $j$-th triangle of the surface of the $k$-th girl, its barycenter $x^k_j$ and the normal vector to the surface in $x^k_j$, $\tau^k_j$, are calculated. Then, the contour of body of a child is associated with the vector field $\sum_{j=1}^{4668} K(x^k_j, \cdot)(\tau^k_j)$ in $H_K(D, \mathbb{R}^3)$ where $D = [-472.73, 487.26] \times [-824.72, 735.28] \times [-156.69, 203.30]$ mm.

Following the same procedure as with the experimental databases, the $k$-means algorithm with $k=2$ is applied to each group corresponding to each height range (1190–1250, 1250–1310, 1310–1370, 1370–1430 mm) within the sample. By using
Table 4  First sizing system proposed for heights between 1190 and 1430 mm

| Size         | Height | Chest length | Waist length | Hip length | Group size |
|--------------|--------|--------------|--------------|------------|------------|
| T1 girl lower (1190–1250) | 1209   | 604          | 538          | 656        | 29         |
| T2 girl upper (1190–1250) | 1227   | 670          | 610          | 739        | 15         |
| T3 girl lower (1250–1310) | 1273   | 643          | 563          | 688        | 20         |
| T4 girl upper (1250–1310) | 1282   | 696.5        | 643.5        | 767        | 31         |
| T5 girl lower (1310–1370) | 1331   | 644          | 564.5        | 701        | 23         |
| T6 girl upper (1310–1370) | 1346   | 733          | 669          | 807        | 32         |
| T7 girl lower (1370–1430) | 1393   | 677          | 586          | 750        | 31         |
| T8 girl upper (1370–1430) | 1410.5 | 797.5        | 722          | 856.5      | 14         |

Fig. 13  Girls who are associated with T1 (first row) and T2 (second row) in the first sizing system

$\lambda_1 = 170.69, \lambda_2 = 178.39, \lambda_3 = 186.98$ and $\lambda_4 = 196.26$, respectively, in each range, we obtain eight groups, which define the first new recommend sizing system. Each size is described using the median values (mm) of the anthropometric measurements of the group, as in Table 4:

Each one of the Figs. 13, 14, 15 and 16 shows part of a group of girls who belong to the same height range. The bodies of the first row of each image are associated with the corresponding lower size, and the second row with the corresponding upper size in the new sizing system.

Moreover, in Fig. 17 it is possible to compare some anthropometric measurements of the two groups obtained within the 1190–1250 mm height range. Something similar occurs in the height.

In addition, MDS has been applied on the distance matrices computed in each range of height to observe the groups obtained (Fig. 18).

To reduce the number of sizes in the previous model, a second model is proposed, which is probably cheaper for the clothing companies. It consists of considering all members of the subsample together and applying $k$-means algorithm to divide the sample into different groups according to the shape and height of the children.
As mentioned above, the question of which number $k$ of clusters to choose is a difficult problems in data clustering. Several methods have been proposed and used in the literature to make this decision (Kaufman and Rousseeuw 1990; Jain 2010).
In our application we combine the “elbow criterion” with a goodness of clustering measure: the silhouette.

The idea of the “elbow criterion” is based on plotting the objective function of Eq. 21 against the number of clusters. The first clusters greatly decrease the objective function, but at some point the marginal decrement will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the “elbow criterion”.

The silhouette of an object is a measure of how close is the object in the neighboring clusters compared with data within its cluster. We use the silhouette because it is a standard measure to validate the clustering structure that combine both cohesion and separation. Given an object \( \phi_i \), let \( a(i) \) be its average 3D dissimilarity with all other objects within its cluster and let \( b(i) \) be the lowest such value for all the other clusters of which \( \phi_i \) is not a member. Then the silhouette of \( \phi_i \) is:

\[
s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)},
\]
Fig. 18 MDS applied on the distance matrices computed in each height range (first sizing system). a Height range 1190–1250 mm, b height range 1250–1310 mm, c height range 1310–1370 mm, d height range 1370–1430 mm

The silhouette is a relative measure that ranges from -1 to 1. A silhouette close to 1 implies the datum is in an appropriate cluster, while a silhouette close to -1 implies the datum is in the wrong cluster. The average $s$ of the silhouette gives us a measure of the goodness of a clustering:

$$s = \frac{s(i)}{m}.$$ 

In Fig. 19 we can see the plot of the objective function (a) and the silhouette (b) against the number of clusters. Taking into account both criteria, we chose to apply the k-means algorithm with $k = 5$ and $\lambda = 183.43$. A partition of the sample is therefore obtained according to the height and size of the children. Figure 20 shows the MDS result applied on the distance matrix computed and Fig. 21 shows the box diagrams of some anthropometric measurements within each group generated.

Moreover, we define the new sizing system based on these groups using the median of the anthropometric measurements from each group, as shown in Table 5:
Fig. 19 Study of the “elbow criterion” in (a) and silhouette for different values of k in (b)

Fig. 20 MDS applied on the distance matrix computed (second sizing system)

Figure 22 shows part of a group of girls. Each one of the rows in the image corresponds to one of the sizes in the second sizing system, (from T1 in the first row, to T5 in the last row). Moreover, the same red segment appears next to each row, to appreciate better the height differences between rows.

8 Discussion

In this paper we have proposed an approach that represents a novel method in terms of using the current-based approximation to shape and size analysis in a clustering procedure and it has been applied in order to define a more efficient children’s sizing system. The data are transformed on elements in an RKHS space and the well-known k-
Fig. 21 Height (a), waist length (b), hip length (c) and chest length (d) from the 5 groups obtained (1190–1430 mm)

Table 5 Second sizing system proposed for heights between 1190 and 1430 mm

| Size  | Height | Chest length | Waist length | Hip length | Group size |
|-------|--------|--------------|--------------|------------|------------|
| T1 girl | 1241   | 610          | 540          | 661        | 57         |
| T2 girl | 1259   | 678          | 622          | 754        | 39         |
| T3 girl | 1362   | 660          | 571.5        | 723.5      | 56         |
| T4 girl | 1361.5 | 747.5        | 673          | 814        | 34         |
| T5 girl | 1417   | 832          | 767          | 903        | 9          |

means clustering algorithm has been adapted to it. An experimental study with simple synthetic objects was successfully conducted to validate the procedure.

We have proposed two ways of defining an efficient sizing system. First, we segmented the data set using height, which is currently the most widely used method. We then applied the k-means algorithm with the number of sizes established within each class as $k = 2$. In this way, the first segmentation provides a first easy input to choose
Fig. 22  Girls classified using the second sizing system. (from T1 in the first row, to T5 in the last row)

the size, while the resulting clusters obtained optimize shape classification within each initial input. This first classification would provide a large percentage of accommodation but perhaps an excessively large number of sizes. To reduce the number of sizes, in the second system proposed, the \textit{k-means} algorithm is applied to all members of the subsamples and goodness of clustering criteria were applied to choose the optimal number of groups. In both cases, the clustering results have been described in terms of anthropometric dimensions within each group.

\textbf{Acknowledgements}  This paper has been partially supported by the Spanish Ministry of Science and Innovation Project DPI2013 – 47279 – C2 – 1 – R. We would also like to thank the Valencian Institute of Biomechanics for providing us with the data set. S.
References

Allen B, Curless B, Popović Z (2003) The space of human body shapes: reconstruction and parameterization from range scans. ACM Trans Graph TOG 22:587–594
Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68:337–404
Baek SY, Lee K (2012) Parametric human body shape modeling framework for human-centered product design. Comput Aided Des 44(1):56–67
Bauer M, Harms P, Michor PW (2011) Sobolev metrics on shape space of surfaces. J Geom Mech 3(4):389–438
Bock HH (2007) Clustering methods: a history of k-means algorithms. In: Brito P, Bertrand P, Cucumel G, de Carvalho F (eds) Selected contributions in data analysis and classification. Springer, Berlin, pp 161–172
Caponnetto A, Micchelli CA, Pontil M, Ying Y (2008) Universal multi-task kernels. J Mach Learn Res 9:1615–1646
Carmeli C, De Vito E, Toigo A (2006) Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. Anal Appl 4(04):377–408
Chung M, Lina H, Wang MJJ (2007) The development of sizing systems for taiwanese elementary- and high-school students. Int J Ind Ergon 37:707–716
Conway JB (2013) A course in functional analysis, vol 96. Springer, Science & Business Media
Cox TF, Cox MA (2000) Multidimensional scaling. CRC press
Durrleman S (2010) Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution. PhD thesis, Université Nice Sophia Antipolis
Durrleman S, Pennec X, Trouvé A, Ayache N (2009) Statistical models of sets of curves and surfaces based on currents. Med Image Anal 13(5):793–808
European Committee for Standardization (2002) European Standard EN 13402-2: Size system of clothing. http://esearch.cen.eu/esearch/Details.aspx?id=5430955
Glavins J (2005) Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l’anatomie numérique. PhD thesis, Université Paris 13. http://cis.jhu.edu/joan/TheseGlauens.pdf. Accessed Sept 2005
Glaunès JA, Joshi S (2006) Template estimation form unlabeled point set data and surfaces for computational anatomy. In: 1st MICCAI workshop on mathematical foundations of computational anatomy: geometrical, statistical and registration methods for modeling biological shape variability
Gual-Arnau X, Herold-García S, Simó A (2015) Apparel sizing using trimmed PAM and OWA operators. Expert Syst Appl 42:10512–10520
Huang H, Wang F, Guibas L (2014) Functional map networks for analyzing and exploring large shape collections. ACM Trans Graph 33(4):1–11
Ibáñez MV, Vinué G, Alemany S, Simó A, Epifanio I, Domingo J, Ayala G (2012) Apparel sizing using trimmed PAM and OWA operators. Expert Syst Appl 29:10512–10520
Jain A, Thurik J, Seidel HP, Theobalt C (2010) Moviereshape: tracking and reshaping of humans in videos. ACM Trans Graph TOG 29:148
Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recognit Lett 31:651–666
Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley
Lang S (1995) Differential and Riemannian manifolds. Springer, New York
Lloyd SP (1957) Least squares quantization in pcm. Bell telephone labs memorandum, Murray Hill, nj. Reprinted in IEEE trans information theory IT-28 (1982) vol 2. pp 129–137
MATLAB (2014) version 8.4.0 (R2014b). The MathWorks Inc., Natick
Micchelli C, Pontil M (2005) On learning vector-valued functions. Neural Comput 17(1):177–204
Morgan F (2008) Geometric measure theory: a beginner’s guide. Academic Press, Cambridge
Nazeer KAA, Sebastian MP (2009) Improving the accuracy and efficiency of the k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24(7):881–892
S. Barahona et al.
Ovsjanikov M, Ben-Chen M, Solomon J, Butscher A, Guibas L (2012) Functional maps: a flexible representation of maps between shapes. ACM Trans Graph 31(4):1–11
Pennec X (2006) Intrinsic statistics on riemannian manifolds: basic tools for geometric measurements. J Math Imaging Vis 25:127–154
Pishchulin L, Wuhrer S, Helten T, Theobalt C, Schiele B (2015) Building statistical shape spaces for 3d human modeling. arXiv:1503.05860
Quang MH, Kang SH, Le TM (2010) Image and video colorization using vector-valued reproducing kernel hilbert spaces. J Math Imaging Vis 37(1):49–65
Steinhaus H (1956) Sur la division des corps matériels en parties. Bull Acad Pol Sci IV(12):801–804
Vaillant M, Glaunès J (2005) Surface matching via currents. Biennial international conference on information processing in medical imaging. Springer, Berlin, Heidelberg, pp 381–392
Vinué G, Simó A, Alemany S (2016) The k-means algorithm for 3d shapes with an application to apparel design. Adv Data Anal Classif 10(1):103–132