3D Shape Sequence of Human Comparison and Classification using Current and Varifolds

Emery Pierson, Mohamed Daoudi, Sylvain Arguillere

3D Shape Sequence of Human Comparison and Classification using Current and Varifolds

Emery Pierson[1], Mohamed Daoudi[1,2], and Sylvain Arguillere[3]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
emery.person@univ-lille.fr
[2] IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems
mohamed.daoudi@imt-nord-europe.fr
[3] Univ. Lille, CNRS, UMR 8524 Laboratoire Paul Painlevé, Lille, F-59000, France
sylvain.arguillere@univ-lille.fr

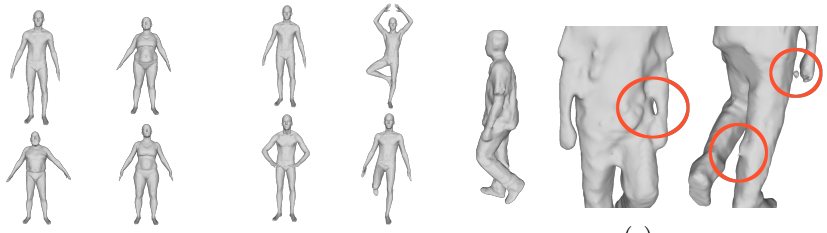
Abstract. In this paper we address the task of the comparison and the classification of 3D shape sequences of human. The non-linear dynamics of the human motion and the changing of the surface parametrization over the time make this task very challenging. To tackle this issue, we propose to embed the 3D shape sequences in an infinite dimensional space, the space of varifolds, endowed with an inner product that comes from a given positive definite kernel. More specifically, our approach involves two steps: 1) the surfaces are represented as varifolds, this representation induces metrics equivariant to rigid motions and invariant to parametrization; 2) the sequences of 3D shapes are represented by Gram matrices derived from their infinite dimensional Hankel matrices. The problem of comparison of two 3D sequences of human is formulated as a comparison of two Gram-Hankel matrices. Extensive experiments on CVSSP3D and Dyna datasets show that our method is competitive with state-of-the-art in 3D human sequence motion retrieval. Code for the experiments is available at https://github.com/CRISTAL-3DSAM/HumanComparisonVarifolds

Keywords: 3D Shape Sequence · Varifold · 3D Shape Comparison · Hankel matrix

1 Introduction

Understanding 3D human shape and motion has many important applications, such as ergonomic design of products, rapid modeling of realistic human characters for virtual worlds, and an early detection of abnormality in predictive clinical analysis. Recently, 3D human data has become highly available as a result of the availability of huge MoCap (Motion Capture) datasets [1,4] along with the evolution of 3D human body representation [25] leaded to the availability of huge artificial human body datasets [26,34]. In the meantime, evolutions in 4D technology for capturing moving shapes lead to paradigms with new multi-view and 4D scan acquisition systems that enable now full 4D models of human shapes that include geometric, motion and appearance information [36,10,31,16].

The first difficulty in analyzing shapes of 3D human comes from noise, variability in pose and articulation, arbitrary mesh parameterizations during data collection, and shape variability within and across shape classes. Some examples of 3D human highlighting these issues are illustrated in Figure 1. In particular, the metrics and representations should have certain invariances or robustness to the above-mentioned variability. Recently, Kaltenmark *et al.* [22] have proposed a general framework for 2D and 3D shape similarity measures, invariant to parametrization and equivariant to rigid transformations. More recently, Bauer *et al.* [7], adopted the varifold fidelity metric as a regularizer for the problem of reparameterization in the framework of elastic shape matching using the SRNF [19] representation. Motivated by the progress of using varifolds and current in shape analysis, we propose to compare 3D surface of human shapes by comparing their varifolds.



(a) (b) (c)

Fig. 1: **Different challenges of 3D Human Sequence Comparison:** (a) Shape variability within and across shape classes, (b) variability in pose and articulation, (c) noisy and arbitrary mesh parameterizations (topological noise, vertex noise and disconnected components).

As a second difficulty, it is critical to identify precise mathematical representations of underlying shapes and then impose efficient dynamical models on representation spaces that capture the essential variability in shape evolutions. In addition to the nonlinearity of shape spaces, one expects nonlinearity in temporal evolutions that makes the inference process difficult. In this paper, we propose to use Gram matrices derived from Hankels matrices to represent the dynamic of human motion.

In our approach, as illustrated in Figure 2, we propose to embed the human shape space $\mathcal{H}$ in an infinite dimensional Hilbert space with inner product corresponding to a positive definite kernel $\langle .,. \rangle_V$ inspired by the varifold framework. Using this kernel product we are able to compute the Gram matrix relative to a motion. Each of this Gram matrix is transformed to Gram-Hankel matrix of fixed size r.

In summary, the main contributions of this article are: ***(i)*** We represent 3D human surfaces as varifold. This representation is equivariant to rotation

and invariant to the parametrization. This representation allows us to define an inner product between two 3D surfaces represented by varifolds. ***(ii)*** It is the first use of the space of varifolds in human shape analysis. The framework does not assume that the correspondences between the surfaces are given. ***(iii)*** We represent 4D surfaces by Hankel matrices. This key contribution enables the use of standard computational tools based on the inner product defined between two varifolds. The dynamic information of a sequence of 3D human shape is encapsulated in Hankel matrices and we propose to compare sequences by using the distance between the resulting Gram-Hankel matrices; ***(iv)*** The experiments results show that the proposed approach improves 3D human motion retrieval state-of-the-art and it is robust to noise.

2 Related Work

2.1 3D Human Shape Comparison

The main difficulty in comparing human shapes of such surfaces is that there is no preferred parameterization that can be used for registering and comparing features across surfaces. Since the shape of a surface is invariant to its parameterization, one would like an approach that yields the same result irrespective of the parameterization. The linear blending approaches [17,3,25] offer a good representation for human shape, along with a model of human deformations while being able to distinguish shape and pose deformations. However these methods need additional information on the raw scans such as MoCap markers [17,3], gender of the body, or additional texture information [25,10] to retrieve such representations. Recently, deep learning approaches [35,6,43] propose human bodies latent spaces that share common properties with linear blending models. However, they require training data with the same mesh parameterization and are sensitive to noise. Moreover, most current techniques treat shape and motion independently, with devoted techniques for either shape or motion in isolation.

Kurtek *et al.* [23] and Tumpach *et al.* [37] propose the quotient of the space of embeddings of a fixed surface S into $\mathbb{R}^3$ by the action of the orientation-preserving diffeomorphisms of S and the group of Euclidean transformations, and provide this quotient with the structure of an infinite-dimensional manifold. The shapes are compared using a Riemannian metric on a *pre-shape space* $\mathcal{F}$ consisting of embeddings or immersions of a model manifold into the 3D Euclidean space $\mathbb{R}^3$. Two embeddings correspond to the same shape in $\mathbb{R}^3$ if and only if they differ by an element of a shape-preserving transformation group. However the use of these approaches on human shape analysis assume a spherical parameterization of the surfaces. Pierson *et al.* [29] propose a Riemannian approach for human shape analysis. This approach provides encouraging results, but it requires the meshes to be registered to a template. Recently, the framework of varifolds have been presented for application to shape matching. Charon *et al.* [13] generalize the framework of currents, which defines a restricted type of geometric measure on surfaces, by the varifolds framework representing surfaces

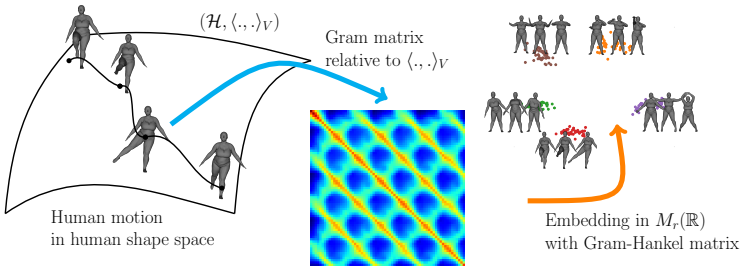


Fig. 2: **Overview of our method**. We embed the human shape space $\mathcal{H}$ in an infinite dimensional Hilbert space with inner product from a positive definite kernel $\langle ., . \rangle_V$ inspired by varifold framework. Using this kernel product we are able to compute Gram matrix relative to a motion. Each of this Gram matrix is transformed to Gram Hankel matrix of size r. The Frobenius distance in $M_r(\mathbb{R})$ is used to retrieve similar 3D sequences.

as a measure on $\mathbb{R}^3 \times \mathbb{S}^2$. The proposed varifolds representation is parameterization invariant, and does not need additional information on raw scans. Inspired by these recent results, we will demonstrate the first use of this mathematical theory in 3D human shape comparison.

2.2 3D Human Sequence Comparison

A general approach adopted when comparing 3D sequences is the extension of static shape descriptors such as 3D shape distribution, Spin Image, and spherical harmonics to include temporal motion information [18,40,30]. While these approaches require the extraction of shape descriptors, our approach does not need a 3D shape feature extraction. It is based on the comparison of surface varifolds within a sequence. In addition, the comparison of 3D sequences require an alignment of the sequences. The Dynamic Time Warping (DTW) algorithm was defined to match temporally distorted time series, by finding an optimal warping path between time series. It has been used for several computer vision applications [8,21] and alignment of 3D human sequences [40,30]. However, DTW does not define a proper distance (no triangle inequality). In addition, a temporal filtering is often required for the alignment of noisy meshes [32]. Our approach enables the comparison of sequences of different temporal duration, does not need any alignment of sequences and is robust to noisy data. We model a sequence of 3D mesh as a dynamical system. The parameters of the dynamical system are embedded in our Hankel matrix-based representation. Hankel matrices have already been adopted successfully for skeleton action recognition in [42]. As we do not have finite dimensional features to build such matrix numerically, we define a novel Gram-Hankel matrix, based on the kernel product defined from surface varifold. This matrix is able to model the temporal dynamics of the 3D meshes.

3 Proposed Method

3.1 Comparing 3D Shapes using Geometric Measures

The varifolds framework is a geometry theory used to solve famous differential geometry problems such as the Plateau's Problem. We invite the interested reader to read an introduction of the theory in [2]. We focus here on the work of Charon *et al.* [13], followed by Kaltenmark *et al.* [22] who proposed to use the varifolds framework for discretized curves and surfaces. They designed a fidelity metric using the varifold representation. This fidelity metric is proposed for 2D artificial contour retrieval, and used in 3D diffeomorphic registration in the Large deformation diffeomorphic metric mapping (LDDMM) framework. To our knowledge our work is the first use of such representation for the analysis of human shape. As far as the authors are aware, our work is the first use of such representation for the analysis of human shape. It is also the first use of the space of varifolds purely for itself, as an efficient way to perform direct computations on shapes.

A *varifold* is a measure μ on $\mathbb{R}^3 \times \mathbb{S}^2$. The integral of a function $f : \mathbb{R}^3 \times \mathbb{S}^2 \to \mathbb{R}$ with respect to such a measure is denoted $\int_{\mathbb{R}^3 \times \mathbb{S}^2} f d\mu$. Given S a smooth compact surface with outer normal unit vector field $x \mapsto n(x)$, the core idea of [13] is to represent S as a varifold. This is done in practice through the formula $\int_{\mathbb{R}^3 \times \mathbb{S}^2} f dS = \int_S f(x, n(x)) dA(x)$, with $dA(x)$ the surface area measure of S at x. Now, given a triangulated surface M of a 3D human shape with triangle faces $T_1, ..., T_m$, when the triangles are sufficiently small, each triangular face T_i is represented as an atomic measure $a_i \delta_{c_i, n_i}$ where c_i is the barycenter of the triangulated face, n_i the oriented normal of the face, a_i its area, and δ representing the dirac mass. The varifold representation of the total shape M is simply given by the sum of all these measures: $M = \sum_{i=1}^{m} a_i \delta_{c_i, n_i}$. To illustrate, integrating a function f on $\mathbb{R}^3 \times \mathbb{S}^2$, with respect to M yields $\int_{\mathbb{R}^3 \times \mathbb{S}^2} f dM = \sum_{i=1}^{m} a_i f(c_i, n_i)$.

A varifold μ can be converted into a function Φ_μ on $\mathbb{R}^3 \times \mathbb{S}^2$ using a *reproducing kernel* that comes from the product of two positive definite kernels: $k_{pos} : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, and $k_{or} : \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}$. We just define $\Phi_\mu(x, v) = \int_{\mathbb{R}^3 \times \mathbb{S}^2} k_{pos}(y, x) k_{or}(w, v) d\mu(y, w)$. For a triangulated surface M, we get $\Phi_M(x, v) = \sum_{i=1}^{m} a_i k_{pos}(c_i, x) k_{or}(n_i, v)$.

One obtains a Hilbert product between any two varifolds μ, ν as follows : $\langle \mu, \nu \rangle_V = \langle \Phi_\mu, \Phi_\nu \rangle_V = \int \Phi_\mu d\nu = \int \Phi_\nu d\mu$, so that

$$\langle \mu, \nu \rangle_V = \iint k_{pos}(x, y) k_{or}(v, w) d\mu(x, v) d\nu(y, w).$$

We deduce the explicit expression for that product between two triangulated 3D shapes M and N :

$$\langle M, N \rangle_V = \langle \Phi_M, \Phi_N \rangle_V = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i^M a_j^N k_{pos}(c_i^M, c_j^N) k_{or}(n_i^M, n_j^N) \qquad (1)$$

Where m, n are the number of faces of M and N. The continuous version of this product presented in [13] is parametrization invariant.

An important part of such a product is that it can be made equivariant to rigid transformation by carefully choosing the kernels. First we define how to apply such a deformation on a varifold. Given a rotation $R \in SO(3)$ of $\mathbb{R}^3$ and a vector $T \in \mathbb{R}^3$, the rigid transformation $\phi : x \mapsto Rx + T$ yields the push-forward transformation $\mu \mapsto \phi_{\#}\mu$ through $\int f d\phi_{\#}\mu = \int f(Rx+T, Rv)d\mu$ on the space of varifolds. For a triangulated surface M, $\phi_{\#}M$ is just $\phi(M)$, the surface obtained by applying the rigid motion ϕ to the surface M itself. We have the following important result :

Theorem 1. *If we define the positive definite kernels as following:*

$$\begin{aligned} k_{pos}(x, y) &= \rho(||x - y||), \quad x, y \in \mathbb{R}^3, \\ k_{or}(v, w) &= \quad \gamma(v.w), \quad v, w \in \mathbb{S}^2, \end{aligned}$$

then for any two varifolds μ, ν, and any rigid motion ϕ on $\mathbb{R}^3$, we have

$$\langle \phi_{\#}\mu, \phi_{\#}\nu \rangle_V = \langle \mu, \nu \rangle_V.$$

This result means that given a rigid motion ϕ, $\langle \phi(M), \phi(N) \rangle_V = \langle M, N \rangle_V$.

The kernel k_{pos} is usually chosen as the Gaussian kernel $k_{pos} = e^{-\frac{||x-y||^2}{\sigma^2}}$, with the scale parameter σ needed to be tuned for each application.

Kaltenmark *et al.* [22] proposed several function for the γ function of the spherical kernel. In this paper we retained the following functions: $\gamma(u) = u$ – *currents*, $\gamma(u) = e^{2u/\sigma_o^2}$ – *oriented varifolds*, and we propose $\gamma(u) = |u|$ – *absolute varifolds*. For such kernels, two surface varifolds M, N with "similar" support (for example, if M is a reparametrization of N, or if they represent two human shapes with the same pose but different body types) will have relatively small distance in the space of varifolds, so that $\langle M, N \rangle_V^2 \simeq \langle M, M \rangle_V \langle N, N \rangle_V$, that is, they are almost co-linear. On the other hand, surface varifolds with very distant support will be almost orthogonal ($\langle M, N \rangle_V \simeq 0$) because of the Gaussian term in k_{pos}. Obviously, shapes that have some parts that almost overlap while others are far away will be in-between. Combined with its rotational invariance, this leads us to believe that the kernel product can be used to differentiate between poses and motions independently of body types.

3.2 Comparing 3D Human Sequences

We need a way to compare sequences of $3D$ shapes $M_1, ..., M_T$, with T possibly differing between sequences. For this, we use the kernel product $\langle ., . \rangle_V$ as a similarity metric. Thanks to the reproducing property of positive definite kernels [5], it defines a reproducing kernel Hilbert space $\mathcal{H}$ (RKHS) which is an (infinite dimensional) Euclidean space endowed with an inner product corresponding to the kernel product, as described in the previous section. Any shape M has a corresponding representative $\Phi_M : \mathbb{R}^3 \times \mathbb{S}^2 \to \mathbb{R}$ in this space, such that $\langle \Phi_M, \Phi_N \rangle_V = \langle M, N \rangle_V$ (Figure 3).

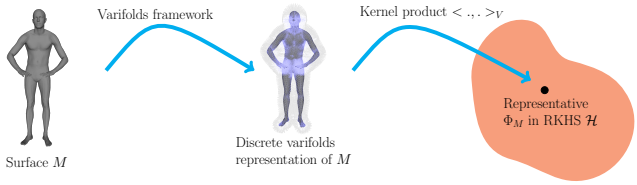


Fig. 3: An overview of varifolds framework. First, a mesh M is transformed into its corresponding varifold representation. Then, the kernel product defined in Equation (1), transforms it into a representant Φ_M living in the Hilbert space $\mathcal{H}$ of this varifold.

Modeling dynamics of temporal sequences. Thanks to the varifolds representation and the kernel product $\langle .,. \rangle$, the temporal sequence $M_1, ..., M_t$ corresponding to a motion in the human shape space can be seen as a temporal sequence $\Phi_{M_1}, ..., \Phi_{M_T}$ in the RKHS $\mathcal{H}$. Plus, since the varifold kernel is equivariant to rigid transformations, the product of two shapes within a sequence is invariant to any rigid transformation *applied to the full motion*. The Gram matrix $J_{ij} = \langle \Phi_{M_i}, \Phi_{M_j} \rangle$, which is a rigid transformation invariant matrix, would be a natural representant of the motion. However, its size vary with the length T of the sequence. Inspired by Auto-Regressive (AR) models of complexity k defined by $\Phi_{M_t} = \sum_{i=1}^{k} \alpha_i \Phi_{M_{t-i}}$, several representations [33,38] have been proposed for dynamical systems modeling. Hankel matrices [24] are one of the possible representations. The Hankel matrix of size r, s corresponding to our time series $\Phi_{M_1}, ..., \Phi_{M_T}$ is defined as:

$$\mathbf{H}_t^{r,s} = \begin{pmatrix} \Phi_{M_1} & \Phi_{M_2} & \Phi_{M_3} & ... & \Phi_{M_s} \\ \Phi_{M_2} & \Phi_{M_3} & \Phi_{M_4} & ... & \Phi_{M_{s+1}} \\ ... & ... & ... & ... & ... \\ \Phi_{M_r} & \Phi_{M_{r+1}} & \Phi_{M_{r+2}} & ... & \Phi_{M_{r+s}} \end{pmatrix} \tag{2}$$

The rank of such matrix is usually, under certain conditions, the complexity k of the dynamical system of the sequence. The comparison of two time series therefore become a comparison of high dimensional matrices.

It is not straightforward to use those matrices since our shape representatives live in infinite dimensional space. A first idea would be to think about the Nystrom reduction method [41] to build an explicit finite dimensional representation for Φ_M, but this would involve intensive computations. Another possibility is to think about the Gram matrix $\mathbf{H}\mathbf{H}^T$ derived from the Hankel matrix $\mathbf{H}$ [42,24]. We cannot directly derive the same kind of matrices since our representatives live in an infinite dimensional space. The Gram matrix of the motion, J, however, preserves the linear relationships of the AR model. We therefore derive the following matrix:

Definition 1. *The Gram-Hankel matrix of size r, $G \in M_r(\mathbb{R})$ of the sequence $\Phi_{M_1}, ..., \Phi_{M_T}$ is defined as:*

$$\mathbf{G}_{ij} = \sum_{k=1}^{T-r} \langle \Phi_{M_{i+k}}, \Phi_{M_{j+k}} \rangle = \sum_{k=1}^{T-r} \langle M_{i+k}, M_{j+k} \rangle_V \tag{3}$$

We normalize $\mathbf{G}$ relatively to the Frobenius norm, following recommended practices [42]. This matrix is the sum of the diagonal blocks B_l^r of size r of the Gram matrix of the sequence pairwise inner products. A possible way of interpreting what encodes a single block B_l^r of size r when $r \geq k$ is to follow the idea of [21] the polar decomposition of the coordinate matrix of $\Phi_{M_l}, ...\Phi_{M_{r+l}}$. This coordinate matrix exists in the space $span(\Phi_{M_0}, ...\Phi_{M_k})$ under to the AR model hypothesis (any Φ_{M_j} is a linear combination of the first k Φ_{M_i}), and can be factorized into the product $U_l R_l$, where U_l is an orthonormal $r \times k$ matrix, and R_l an SPD matrix of size k. The matrix R_l is the covariance (multiplied by r^2) of $\Phi_{M_l}, ...\Phi_{M_{r+l}}$ in $span(\Phi_{M_0}, ...\Phi_{M_k})$, and it encodes in some way its shape in this space. An illustration of such encoding is given in Figure 4. For three motions from CVSSP3D dataset, we compute the varifold distances Equation (1) between all samples of the motion. We then used Multidimensional Scaling (MDS) [15] to visualize them in a 2D space. We display the ellipse associated to the covariance of each motion. We see that the one associated to jump in place (blue) motion is distinguishable from the ones associated to walk motions (red and green).

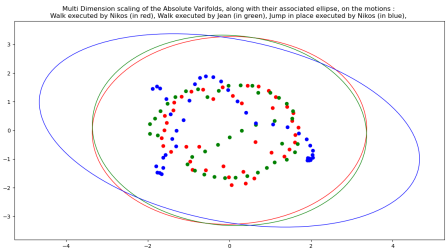


Fig. 4: MDS illustration of three motions of CVSSP3D dataset, along with ellipse associated to their covariance.

The Gram matrix block B_l^r is written as $B_l^r = U_l R_l^2 U_l^T$ and contains such information. Searching for the complexity k of the AR model would be sensitive to errors, and computing the associated R_l for comparisons with an SPD metric would be time consuming in our case. We thus preferred the rather simpler Gram-Hankel matrix, that cancels possible noise in single blocks when summing them. Finally, using the Frobenius distance $d(\mathbf{G_i}, \mathbf{G_j}) = ||\mathbf{G_i} - \mathbf{G_j}||_F$ where G_i and G_j are two Gram-Hankel matrices, to compare two motions lead us to rather good results. The blocks of size r are expressive enough when $r \geq k$, and taking

their sum will ensure us to cancel possible noise added to a single block. The degenerate nature of G, does not allow for efficient use of SPD metrics such as the Log-Euclidean Riemannian Metric (LERM) on the G_i (more details are available in the supplementary material). With this approach, the comparison of two human motions is formulated as the comparison of two symmetric positive semi-definite matrices.

Proposition 1. *The Gram-Hankel matrix G associated to a motion $M_1, ..., M_T$ defined by Equation (3) has the following properties:*

1. *It is invariant to parameterization (property of the kernel product).*
2. *It is invariant to rigid transformation applied to a motion.*

Normalizations As the definition of the kernel shows the method is not invariant to scale, we normalize the inner products as following: $\left\langle \frac{M_i}{||M_i||_V}, \frac{M_j}{||M_j||_V} \right\rangle_V$.

While our method is translation invariant, the use of the Gaussian kernel implies that the product will be near 0 when the human shapes are at long range. To avoid this, we translate the surface M with triangles $T_1, ..., T_m$ by its centroid $c_M = \frac{\sum_{i=1}^m a_i c_i}{\sum_{i=1}^m a_i}$, where c_i and a_i correspond to the center and area of triangle T_i. We apply $M \mapsto M - c_M$ before computing the products.

4 Experiments

Computing varifold kernel products can often be time consuming, due to the quadratic cost in memory and time in terms of vertex number for computing $\langle M, N \rangle_V$. However, the recent library Keops [12], designed specifically for kernel operations proposes efficient implementations with no memory overflow, reducing time computation by two orders of magnitudes. We used those implementations with the Pytorch backend on a computer setup with Intel(R) Xeon(R) Bronze 3204 CPU @ 1.90GHz, and a Nvidia Quadro RTX 4000 8GB GPU.

4.1 Evaluation setup

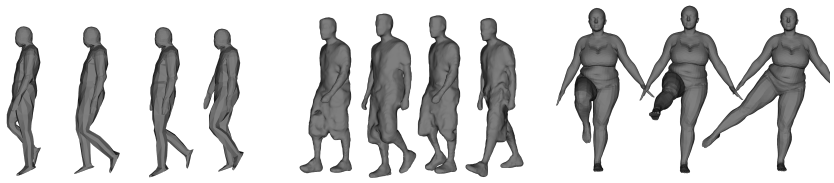
In order to measure the performance in motion retrieval, we use the classical performance metrics used in retrieval: Nearest neighbor (NN), First-tier (FT) and Second-tier (ST) criteria. For each experiment, we take r values ranging from 1 to T_{min} where T_{min} is the minimal sequence length in the dataset. We also take 10 σ values for the Gaussian kernel ranging from 0.001 to 10 in log scale. The score displayed is the best score among all r and σ values. For oriented varifolds, the σ_o of the gamma function is fixed to 0.5 as in [22].

4.2 Datasets

CVSSP3D synthetic dataset [34]. A synthetic model (1290 vertices and 2108 faces) is animated thanks to real motion skeleton data. Fourteen individuals

executed 28 different motions: sneak, walk (slow, fast, turn left/right, circle left/right, cool, cowboy, elderly, tired, macho, march, mickey, sexy, dainty), run (slow, fast, turn right/left, circle left/right), sprint, vogue, faint, rock n'roll, shoot. An example of human motion from this dataset is presented in Figure 5. The frequency of samples is set to 25Hz, with 100 samples per sequences. The maximum computation time of Gram-Hankel matrix is 0.89s.

CVSSP3D real dataset [16]. This dataset contains reconstructions of multi view performances. 8 individuals performed 12 different motions: walk, run, jump, bend, hand wave (interaction between two models), jump in place, sit and stand up, run and fall, walk and sit, run then jump and walk, handshake (interaction between two models), pull. The number of vertices vary between 35000 and 70000. The frequency of samples is also set to 25Hz, and sequence length vary from 50 to 150 (average 109). We keep the 10 individual motions following [40]. An example motion of the dataset is displayed in Figure 5(b). The maximum computation time of Gram-Hankel matrix is 6m30s. The sensitivity of the reconstruction pipeline to the clothes is illustrated by the presence of noise as illustrated in Figure 1. This noise makes this dataset challenging for 3D shape human shape comparison and for 3D human motion retrieval.



(a) (b) (c)

Fig. 5: Example motions from the datasets: (a) Slow walking motion from CVSSP3D synthetic dataset, (b) Walking motion from the CVSSP3D real dataset, (c) Knees motion from the Dyna dataset.

Dyna dataset [31]. This dataset is created from 4D human scans. A human template (6890 vertices) is registered to human body scans sampled at 60 Hz, and sequence length vary from 150 to 1200 (average 323). 10 individuals performed at most 14 different treadmill motions (hips, knees, light hopping stiff, light hopping loose, jiggle on toes, one leg loose, shake arms, chicken wings, punching, shake shoulders, shake hips, jumping jacks, one leg jump, running on spot), which means that the individual only move along the height axis. An example of human motion from Dyna dataset is presented in Figure 5(c). The maximum computation time of Gram-Hankel matrix is 2m30s.

4.3 Motion Retrieval on CVSSP3D Dataset

Comparison with state-of-the-art. We compare our motion retrieval approach to the best features presented in [40,30] and deep learning descriptors: (1) The 3D harmonics descriptor [28][40] is a descriptor based on point cloud repartition in space, (2) Breadths spectrum Q-breadths and Q-shape invariant [30] are presented as 2 fully invariant descriptors derived from convex shape analysis, (3) Aumentado-Armstrong *et al.* [6] propose a human pose latent vector in their Geometrically Disentangled Variational AutoEncoder (GDVAE), (4) Zhou *et al.* [43] propose a human pose latent vector derived from the Neural3DMM [11] mesh autoencoder architecture, and (5) Cosmo *et al.* [14] propose a human pose latent vector in a similar approach as GDVAE, called Latent Interpolation with Metric Priors (LIMP). For the artificial dataset, the optimal σ were fixed to 0.17 for current, 0.17 for absolute varifolds, and 0.02 for oriented varifolds. The optimal r were 97.92 and 90 for current, absolute and oriented varifolds respectively.

We observe the results on the CVSSP3D artificial dataset in Table 1. Only our approach are able to get 100% in all performance metrics. We also observe that it is the only approach able to outperform the LIMP learned approach.

For the real dataset, the optimal σ were fixed to 0.06 for current, 0.17 for absolute varifolds and oriented varifolds. The optimal r were $48, 43$ and 46 for current, absolute and oriented varifolds respectively.

We observe the results on the CVSSP3D real dataset in Table 1. Absolute varifolds approach outperforms by 2.5% the 3D descriptor in terms of NN metric, while being less good for FT and ST. In terms of fully invariant methods, we outperform by 10% the proposed approaches. The absolute varifolds methods is the best of our approach, but we do not observe significant sensitivity between different varifolds. We finally observe that the point cloud descriptors of GDVAE has the lowest performance.

4.4 Motion Retrieval on Dyna Dataset

Comparison with state-of-the-art. No benchmark exists on this dataset, a little has been made on the registrations provided by Dyna. We applied the following methods to extract pose descriptors and made pairwise sequences comparisons using dynamic time warping, in a similar protocol as [40,30], without the temporal filtering use for clothes datasets, since the dataset is not noisy. We compare our approach to descriptor sequences of the following approaches: (1) Areas and Breadths [30] are parameterization and translation invariants derived from convex shapes analysis, (2) The pretrained GDVAE on SURREAL is applied directly on the dataset, (3) the pretrained LIMP VAE on FAUST is applied directly on the dataset, (4) Zhou *et al.* [43] provide pretrained weights on the AMASS dataset [26] for their approach. This dataset shares the same human body parameterization as Dyna, so we can use the pretrained network on Dyna, and (5) The Skinned Multi-Person Linear model (SMPL) body model [25] is a

Representation	Γ inv.	$SO(3)$	Artificial dataset			Real dataset			Dyna dataset		
			NN	FT	ST	NN	FT	ST	NN	FT	ST
Shape Dist. [27][40]	✓	✓	92.1	88.9	97.2	77.5	51.6	65.5	/	/	/
Spin Images [20][40]	✓	✓	**100**	87.1	94.1	77.5	51.6	65.5	/	/	/
3D harmonics [40]	≈	≈	**100**	98.3	99.9	92.5	**72.7**	**86.1**	/	/	/
Breadths spectrum [30]	✓	✓	**100**	99.8	**100**	/	/	/	/	/	/
Shape invariant [30]	✓	✓	82.1	56.8	68.5	/	/	/	/	/	/
Q-Breadths spectrum [30]	≈	✓	/	/	/	80.0	44.8	59.5	/	/	/
Q-shape invariant [30]	≈	✓	/	/	/	82.5	51.3	68.8	/	/	/
Areas [30]	✓	✗	/	/	/	/	/	/	37.2	24.5	35.8
Breadths [30]	✓	✗	/	/	/	/	/	/	50.7	36.2	50.5
Areas & Breadths [30]	✓	✗	/	/	/	/	/	/	50.7	37.2	51.7
GDVAE [6]	✓	✓	**100**	97.6	98.8	38.7	31.6	51.6	18.7	19.6	32.2
Zhou *et al.* [43]	✗	✗	**100**	99.6	99.6	/	/	/	50.0	40.4	57.0
LIMP [14]	✓	✗	**100**	*99.98*	*99.98*	/	/	/	29.1	20.7	33.9
SMPL pose vector [25]	≈	✓	/	/	/	/	/	/	*58.2*	***45.7***	***63.2***
Current	✓	✓	**100**	**100**	**100**	92.5	66.0	78.5	59.0	34.1	50.4
Absolute varifolds	✓	✓	**100**	**100**	**100**	**95.0**	66.6	80.7	**60.4**	40.0	55.9
Oriented varifolds	✓	✓	**100**	**100**	**100**	93.8	65.4	78.2	**60.4**	40.8	55.9

Table 1: Full comparison of motion retrieval approaches. First two columns correspond to group invariance (Γ: reparameterization group, $SO(3)$: rotation group), telling whether or not the required invariance is fullfilled (✓: fully invariant, ≈ : approximately invariant (normalization, supplementary information, ...), ✗: no invariance). Remaining columns correspond to retrieval scores, where the '/' symbol means that there is no result for the method on the given dataset for various reasons, such as unavailable implementation or the method not being adapted for the dataset (for example, in line 470, [41] is based on a given mesh with vertex correspondences and cannot be applied to CVSSP3D real dataset). The results are displayed for CVSSP3D artificial and real datasets, and Dyna datasets. Our method is competitive or better than the approach consisting of combing DTW with any descriptor, while showing all required invariances.

parameterized human body model. We use the pose vector of the body model, computed in [10] using additional information.

For the Dyna dataset, the optimal σ were fixed to 0.02 for current 0.06 for absolute varifolds, and 0.16 for oriented varifolds. The optimal r were $27, 31$ and 72 for current, absolute and oriented varifolds respectively.

As shown in Table 1 the oriented and absolute varifolds is the best by 2 % in terms of NN metric compare to SMPL, and by more than 10 % to other approaches, including the parameterization dependant approach of [43]. The FT and ST performance are however less good than SMPL. This can be explained by its human specific design, along with the costly fitting method, that use additional information (gender, texture videos). Finally, we observe that point cloud neural networks are not suitable for high set of complex motions.

4.5 Qualitative analysis on Dyna dataset

We display in Figure 6 the Nearest Neighbor score confusion matrices for both SMPL and Oriented Varifolds. The confusion matrices for the other datasets are available in the supplementary material. We observe that on Dyna, the difficult cases were *jiggle on toes, shake arms, shake hips and jumping jacks*, corresponding to l3, l6, l9 and l10 in confusion matrix. Our approach is able to classify better these motions than SMPL. In addition, SMPL was not able to retrieve as a Nearest Neighbor, a similar motion to shake arms or shake hips corresponding to l6 and l9. This Figure shows also that our approach retrieves perfectly the knees motion corresponding to l1. The Figure 11 shows some qualitative results of our approach. It illustrates the first tier of a given query on Dyna dataset.

5 Discussion

Effect of the parameters for oriented varifolds on Dyna dataset. We provide in Figure 8, the performance relative to the parameters σ and r, for oriented varifolds on Dyna dataset. We observe that the choice of those parameters is crucial. We also display the performances of oriented varifolds with the 2 normalizations techniques, showing that they both help to obtain the best results. More discussion is provided in the supplementary material.

Limitations. Our approach presents two main limitations: (i) To measure distance between matrices, we have used Euclidean distance, which does not exploit the geometry of the symmetric positive semi-definite matrices manifold, (ii) There is no theoretical limitation to apply this framework to the comparison of other 3D shape sequences (eg. 3D facial expressions, or 3D cortical surfaces evolutions) other than that between body shape. However, in practice one should redefine the hyperparameters (r, σ) of the Kernel (Theorem 1).

6 Conclusion

We presented a novel framework to perform comparison of 3D human shape sequences. We propose a new representation of 3D human shape, equivariant

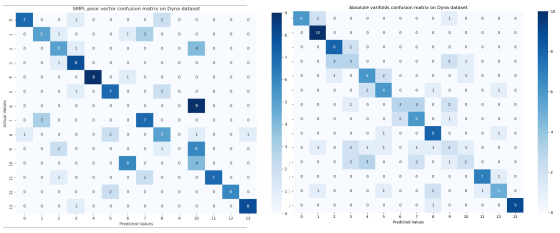
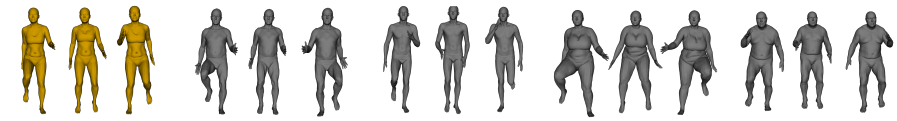
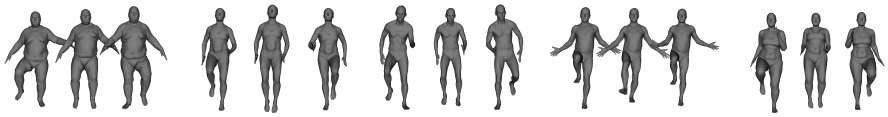


Fig. 6: Confusion matrix of SMPL (left) and Oriented Varifolds (right) on the Dyna dataset.



(a) 50020, running on spot (b) 50026, running on spot (c) 50009, running on spot (d) 50004, running on spot (e) 50007, running on spot



(f) 50002, running on spot (g) 50025, running on spot (h) 50027, running on spot (i) 50026, one leg jump (j) 50022, running on spot

Fig. 7: Qualitative results on Dyna dataset. Given the query corresponding to running on spot motion (a), the first tier result using oriented varifolds are given by (b) – (j).

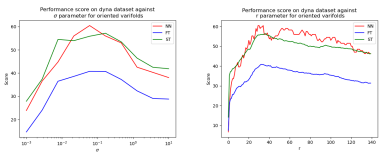


Fig. 8: NN, FT, ST metric relatively to the σ parameters (left) and to the r parameter (right) on Dyna dataset for oriented varifolds.

Centroid	Inner	NN	FT	ST
✗	✗	51.5	34.6	53.4
✗	✓	52.2	33.4	50.5
✓	✗	59.7	40.7	55.8
✓	✓	**60.4**	**40.8**	**55.9**

Table 2: Retrieval performance of the normalizations on Dyna dataset, for oriented varifolds. Both are useful.

to rotation and invariant to parameterization using the varifolds framework. We propose also a new way to represent a human motion by embedding the 3D shape sequences in infinite dimensional space using a kernel positive definite product from varifolds framework. We compared our method to the combination of dynamic time warping and static human pose descriptors. Our experiments on 3 datasets showed that our approach gives competitive or better than state-of-the-art results for 3D human motion retrieval, showing better generalization ability than popular deep learning approaches.

Acknowledgments

This work is supported by the ANR project Human4D ANR-19-CE23-0020 and partially by the Investments for the future program ANR-16-IDEX-0004 ULNE.

References

1. Carnegie Mellon University MoCap Database (2018), http://mocap.cs.cmu.edu/

1
2. Almgren, F.J.: Plateau's problem: an invitation to varifold geometry, vol. 13. American Mathematical Soc. (1966) 5
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416 (2005) 3
4. Aristidou, A., Chrysanthou, Y.: Feature extraction for human motion indexing of acted dance performances. In: Proceedings of the 9th International Conference on Computer Graphics Theory and Applications. pp. 277–287. GRAPP '14, IEEE (2014) 1
5. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American mathematical society **68**(3), 337–404 (1950) 6
6. Aumentado-Armstrong, T., Tsogkas, S., Jepson, A., Dickinson, S.: Geometric disentanglement for generative latent shape models. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8180–8189 (2019) 3, 11, 12, 18, 19
7. Bauer, M., Charon, N., Harms, P., Hsieh, H.W.: A numerical framework for elastic surface matching, comparison, and interpolation. International Journal of Computer Vision pp. 1–20 (2021) 2
8. Ben Amor, B., Su, J., Srivastava, A.: Action recognition using rate-invariant analysis of skeletal shape trajectories. IEEE Trans. on Pattern Analysis and Machine Intelligence **38**(1), 1–13 (2016) 4
9. Bogo, F., Romero, J., Loper, M., Black, M.J.: FAUST: Dataset and evaluation for 3D mesh registration. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3794–3801 (2014) 19, 21
10. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic FAUST: Registering Human Bodies in Motion. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5573–5582. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.591, http://ieeexplore.ieee.org/document/8100074/ 1, 3, 12
11. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Zafeiriou, S., Bronstein, M.M.: Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 7212–7221. IEEE (2019) 11, 18
12. Charlier, B., Feydy, J., Glaunès, J.A., Collin, F.D., Durif, G.: Kernel operations on the gpu, with autodiff, without memory overflows. Journal of Machine Learning Research **22**(74), 1–6 (2021), http://jmlr.org/papers/v22/20-275.html 9
13. Charon, N., Trouvé, A.: The varifold representation of nonoriented shapes for diffeomorphic registration. SIAM journal on Imaging Sciences **6**(4), 2547–2580 (2013) 3, 5, 6
14. Cosmo, L., Norelli, A., Halimi, O., Kimmel, R., Rodolà, E.: Limp: Learning latent shape representations with metric preservation priors. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 19–35. Springer (2020) 11, 12, 18
15. Cox, M.A., Cox, T.F.: Multidimensional scaling. In: Handbook of data visualization, pp. 315–347. Springer (2008) 8
16. Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., Pitas, I.: The i3dpost multi-view and 3d human action/interaction database. In: 2009 Conference for Visual Media Production. pp. 159–168. IEEE (2009) 1, 10

17. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. In: Computer graphics forum. vol. 28, pp. 337–346. Wiley Online Library (2009) 3
18. Huang, P., Hilton, A., Starck, J.: Shape similarity for 3D video sequences of people. International Journal of Computer Vision **89**(2-3), 362–381 (2010) 4
19. Jermyn, I.H., Kurtek, S., Klassen, E., Srivastava, A.: Elastic shape matching of parameterized surfaces using square root normal fields. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 804–817. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 2
20. Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Anal. Mach. Intell. **21**(5), 433–449 (1999) 12
21. Kacem, A., Daoudi, M., Amor, B.B., Berretti, S., Alvarez-Paiva, J.C.: A novel geometric framework on gram matrix trajectories for human behavior understanding. IEEE transactions on pattern analysis and machine intelligence **42**(1), 1–14 (2018) 4, 8
22. Kaltenmark, I., Charlier, B., Charon, N.: A general framework for curve and surface comparison and registration with oriented varifolds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3346–3355 (2017) 2, 5, 6, 9
23. Kurtek, S., Klassen, E., Gore, J.C., Ding, Z., Srivastava, A.: Elastic geodesic paths in shape space of parameterized surfaces. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1717–1730 (2012) 3
24. Li, B., Camps, O.I., Sznaier, M.: Cross-view activity recognition using hankelets. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1362–1369. IEEE (2012) 7
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015) 1, 3, 11, 12, 19
26. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5442–5451 (2019) 1, 11
27. Osada, R., Funkhouser, T.A., Chazelle, B., Dobkin, D.P.: Shape distributions. ACM Trans. Graph. **21**(4), 807–832 (2002) 12
28. Papadakis, P., Pratikakis, I., Theoharis, T., Passalis, G., Perantonis, S.: 3d object retrieval using an efficient and compact hybrid shape descriptor. In: Eurographics Workshop on 3D object retrieval (2008) 11, 18
29. Pierson, E., Daoudi, M., Tumpach, A.B.: A riemannian framework for analysis of human body surface. In: Proceedings of the IEEE IEEE Winter Conf. on Applications of Computer Vision (2022) 3
30. Pierson, E., Paiva, J.C.Á., Daoudi, M.: Projection-based classification of surfaces for 3d human mesh sequence retrieval. Computers & Graphics (2021) 4, 11, 12, 17, 18, 21, 22, 23
31. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. ACM Transactions on Graphics (TOG) **34**(4), 1–14 (2015) 1, 10
32. Slama, R., Wannous, H., Daoudi, M.: 3D human motion analysis framework for shape similarity and retrieval. Image and Vision Computing **32**(2), 131–154 (Feb 2014) 4

33. Slama, R., Wannous, H., Daoudi, M., Srivastava, A.: Accurate 3d action recognition using learning on the grassmann manifold. Pattern Recognition **48**(2), 556–567 (2015) 7
34. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Computer Graphics and Applications **27**(3), 21–31 (2007) 1, 9
35. Tan, Q., Gao, L., Lai, Y.K., Xia, S.: Variational autoencoders for deforming 3d mesh models. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5841–5850 (2018) 3
36. Tsiminaki, V., Franco, J.S., Boyer, E.: High Resolution 3D Shape Texture from Multiple Videos. In: CVPR 2014 - IEEE International Conference on Computer Vision and Pattern Recognition. pp. 1502–1509. IEEE, Columbus, OH, United States (Jun 2014). https://doi.org/10.1109/CVPR.2014.195, https://hal.inria.fr/hal-00977755 1
37. Tumpach, A.B., Drira, H., Daoudi, M., Srivastava, A.: Gauge invariant framework for shape analysis of surfaces. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 46–59 (2016) 3
38. Turaga, P., Veeraraghavan, A., Chellappa, R.: Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008) 7
39. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 4627–4635. IEEE Computer Society (2017) 18
40. Veinidis, C., Danelakis, A., Pratikakis, I., Theoharis, T.: Effective descriptors for human action retrieval from 3d mesh sequences. International Journal of Image and Graphics **19**(03), 1950018 (2019) 4, 10, 11, 12, 17, 18
41. Williams, C., Seeger, M.: Using the nyström method to speed up kernel machines. In: Proceedings of the 14th annual conference on neural information processing systems. pp. 682–688 (2001) 7
42. Zhang, X., Wang, Y., Gou, M., Sznaier, M., Camps, O.: Efficient Temporal Sequence Comparison and Classification Using Gram Matrix Embeddings on a Riemannian Manifold. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4498–4507. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.487, http://ieeexplore.ieee.org/document/7780856/ 4, 7, 8
43. Zhou, K., Bhatnagar, B.L., Pons-Moll, G.: Unsupervised shape and pose disentanglement for 3D meshes. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 341–357. Cham (2020) 3, 11, 12, 18, 19, 21

A Appendix: Comparison with state-of-the-art

In this section, we explain in more details the state-of-the-art methods. Extensive comparison has been made in [40,30] to evaluate the descriptors for human motion retrieval on CVSSP3D dataset. The polygonal curves of those descriptors are filtered with a temporal filtering approach (a mean filter is applied along a temporal window of size K). Finally, the dynamic time warping distance is used for comparing the resulting curves. We compare our motion retrieval approach to the best features presented in those papers, and to several other learned descriptors:

1. The 3D harmonics descriptor [28][40] is a descriptor based on point cloud repartition in space. A 3D shape is first normalized with two variations of PCA. Then, a spherical histogram with different rays is built. The final descriptor is decomposed along spherical harmonics of the obtained with a specific re-weighting for better results. Temporal filtering is proposed in order to deal with the real dataset. We report the results from [40].
2. Breadths spectrum and shape invariant [30] are presented as 2 fully invariant descriptors derived from convex shape analysis. The authors propose to use the breadths of the projection of a shape along each axis spanned by a normal $u \in \mathbb{S}^2$ and to keep the rotation invariant spherical spectrum as a descriptor for human pose. They combine the proposed descriptor with weighted areas of the projection on each plan spanned by u to build a shape invariant. Noise robust version of this descriptor, along with specific temporal filtering named Q-breadths and Q-shape invariant are proposed for the real dataset.
3. Areas, Breadths are the full spherical signals of breadths and weighted areas is proposed to deal with dataset that shows no rotations in [30]. We apply those descriptors, of size 64, along with their concatenation, Areas & Breadths, on Dyna dataset.
4. Aumentado-Armstrong *et al.* [6] propose a variational autoencoder called Geometrically Disentangled VAE (GDVAE). They use PointNet architecture as point cloud encoders and decoders. In the paper, the authors propose to use disentangled intrinsic and extrinsic latent vectors for human shape representation. PointNet encoder is parameterization invariant, but training loss uses the mesh Laplace Beltrami operator which needs a constant parameterization along the training set. Constraints are applied in training to make the network rotation invariant. We report the result of their extrinsic latent vectors (belonging to $\mathbb{R}^{12}$) from [30]. The network was pretrained on the SURREAL dataset [39]. For the CVSSP3D datasets, we report the results from [30].
5. Zhou *et al.* [43] propose a mesh autoencoder based on the Neural3DMM [11] mesh neural network architecture. The network is only applied on human shapes, with the objective to disentangle shape and pose in latent space. The network architecture requires that all input meshes have the same parameterization. We can thus apply it only on the artificial dataset. We report the cross validated results from [30] using the pose latent vectors (belonging to $\mathbb{R}^{112}$) in the human sequence retrieval pipeline. Since the input of the network are the coordinates of the vertices, the approach is not rotation invariant. For the artificial dataset, we report the results from [30].
6. Cosmo *et al.* [14] propose a similar approach as GDVAE, called Latent Interpolation with Metric Priors (LIMP). They use the same type of autoencoder as GDVAE but change the disentanglement constraints with metric prior constraints: a change in extrinsic latent space should only induce change on extrinsic distances of the meshes, while a change in intrinsic latent space should only induce change on intrinsic distances of the meshes. They use Euclidean and geodesic pairwise matrices in their losses to model this constraint, which needs a constant parameterization in the training set. We use

the network pretrained on the FAUST dataset [9]. They do not make any specific training for Euclidean invariance. In order to do motion retrieval, we applied the meshes as input of their available trained network and gathered their extrinsic latent vectors (belonging to $\mathbb{R}^{64}$), and used them in the human sequence retrieval pipeline.

7. Skinned Multi-Person Linear model (SMPL) pose representation. The SMPL body model [25] is a parameterized human body model. A template is deformed (non-rigidly) according to a deformation parameterized by a shape vector. A skeleton is associated to this template and a pose vector, composed of relative rotation of each skeletal joint compared to its parent joint. We convert each joint rotation to quaternion representation as in [43,6] and measure the distance between unit quaternions by $d(q, q') = 1 - |q.q'|$. The SMPL body pose vector contains the pose information of 20 joints, and the rotation of the central joint accounts for the global rotation of the shape, resulting in a $(\mathbb{R}^4)^{20} = \mathbb{R}^{80}$ representation. Due to the construction of the pose vector, this descriptor is rotation invariant. The SMPL parameters were augmented with dynamic soft tissue deformation relative to each motion (called DMPL) and use to transform the original Dyna dataset to the DFAUST dataset, with better correspondance with the scan. They use for this goal much more information such as texture information from body videos, and the shape vector is retrieved using gender information. We prefer comparing on Dyna dataset rather than DFAUST dataset, allowing us to compare faithfully to the SMPL body pose descriptor. In order to build the pose vectors, a costly fitting method is used along each sequence (accounting in minutes for a single shape). The pose vectors for 129 motions of Dyna where the fitting was successful, we added the SMPL Pose vector retrieved using available code https://github.com/vchoutas/smplx/ for the remaining 5 motions.

B Comparison of SPD metrics for Gram-Hankel matrices

This section is dedicated to the comparison between Frobenius and Log Euclidean Riemannian Metric (LERM). The Gram-Hankel matrices are positive semidefinite matrices. Several metrics have been propose to compare positive semidefinite matrices. Table 3 shows the results of the comparison between Log Euclidean Riemannian Metric (LERM) and the Frobenius distance.

$$d_{LERM}(G_1, G_2) = ||\log(G_1) - \log(G_2)||_F,$$

where $\log(G) = P^T \log(\lambda) P$, where $G = P^T \lambda P$ is the eigen decomposition of the symmetric matrix G. We observe that the performance is lower than using the Frobenius metric. This results confirms our choice of using Frobenius than LREM metric.

C Extended discussion on the parameters r and σ

Effect of the sigma parameter. The performance relative to the σ parameter is displayed on the right of Figure 8 in the main paper for oriented varifolds on

Representation	Gram-Hankel distance	Artificial dataset			Real dataset			Dyna dataset		
		NN	FT	ST	NN	FT	ST	NN	FT	ST
Current	Frobenius	100	100	100	92.5	66.0	78.5	59.0	34.1	50.4
	LERM	100	100	100	78.8	55.0	76.6	55.2	35.9	51.4
Absolute varifolds	Frobenius	100	100	100	95.0	66.6	80.7	60.4	40.0	55.9
	LERM	100	100	100	80.0	54.6	73.4	57.5	36.0	50.8
Oriented varifolds	Frobenius	100	100	100	93.8	65.4	78.2	60.4	40.8	55.9
	LERM	100	100	100	86.3	50.0	66.4	57.5	37.0	51.3

Table 3: Motion retrieval results for our approach with Log Euclidean Riemannian Metric (LERM). The results are displayed for CVSSP3D artificial and real datasets, and Dyna datasets

Dyna dataset. We observe first that the choice of σ has a significant impact on performance for NN and in the same time that the optimal σ for the NN is not the same as one for FT and ST, for a loss of around 2% in those metrics, which is less significant than the NN gain.

Effect of the choice of r. The performance relative to the r parameter is displayed on the left of Figure 8 for oriented varifolds on Dyna dataset. We observe first that the choice of r has a significant impact on performance and in the same time that the optimal r for the NN is not the same as one for FT and ST, for a loss of around 5% in those metrics.

Effect of normalizations. We present in Table 2 of the main paper the performances of oriented varifolds with the 2 normalization techniques presented here. The centroid normalization is essential to the good performance of our approach. In the mean time, the inner product normalization always implies significant boost for NN metric, but can induce a (non-significant) loss in ST and FT metrics.

D Qualitative results: Queries on Dyna

Figure 9 shows the results for SMPL, Zhou et al and Areas & Breadths. Although it is a the first tier is better for our approach in two manners: First we observe that there is no confusion between a motion and the motion of the same individual in our approach. Secondly, some drawbacks of the other methods appear: Areas & Breadths are symmetric descriptors and does not make the difference between a punching arm (from down to up) and the two arms that goes up and down when running, and we see a lot of punching motions retrieved (4 out of 6 wrong retrievals). Second, the autoencoder of Zhou *et al.*is not fully disentangled from the identity of the body and a lot of motions from the same identity are retrieved (4 out of 6 wrong retrievals). SMPL gives the best result, as expected from Table 1 of the paper. However, we observe also some sensitivity to the identity of the performer (2 out of 3 wrong retrievals).

Fig. 9: First tier of the query of the paper for Areas & Breadths [30], Zhou *et al.* [43], SMPL [9] and oriented varifolds on the Dyna dataset. The query is in yellow and the results are sorted by closeness to the query using a given approach.

E Qualitative results on CVSSP3D real dataset.

In the CVSSP3D real dataset, clothes worn by the subjects during the acquisition process induce topological and mesh noises (see Figure 1 and Figure 5(b) of the paper). The results on this dataset shows our method robustness to the noise and clothes present in clothed human dataset. The quantitative results in Table 1 (paper) show that our approach is robust to the noise and outperforms state-of-art methods on CVSSP3D real dataset in terms of NN. The confusion matrix of our approach (absolute varifolds) on CVSSP3D real dataset , in Figure 10 shows that our approach performs well on all human motions of the dataset. We display also a query with absolute varifolds, in Figure 11 (same query as the one displayed in [30]). Our approach is able to provide 6 out of the 7 walk motion, showing a slighlty better results compared to [30] (5 out of 7).

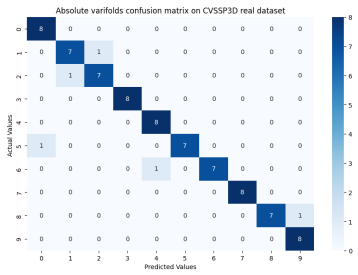


Fig. 10: NN Confusion matrix of absolute varifolds on CVSSP3D real dataset.

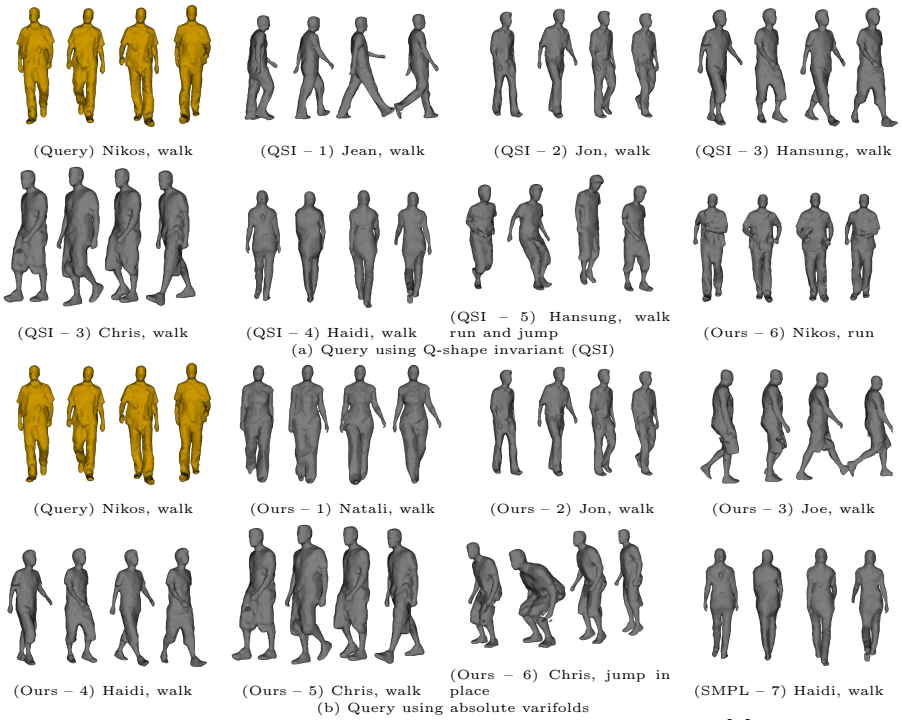


Fig. 11: First tier of the query of the paper for Q-shape invariant [30] and Absolute Varifolds. The query is in yellow and the results are sorted by closeness to the query using a given approach. The first query is directly taken from [30]. The query is in yellow and the results are sorted by closeness to the query using a given approach.