A Geometric ConvNet on 3D Shape Manifold for Gait Recognition

Nadia Hosni
CRISTAL, University of Manouba, Tunisia.
IMT Lille Douai/CRIStAL,
CNRS 9189, University of Lille, France
nadia.hosni@imt-lille-douai.fr

Boulbaba Ben Amor
Inception Institute of Artificial Intelligence, UAE
boulbaba.amor@inceptioniai.org

Abstract

In this work we propose a geometric deep convolutional auto-encoder (DCAE) for the purpose of gait recognition by analyzing time-varying 3D skeletal data. Sequences are viewed as time-parameterized trajectories on the Kendall shape space $\mathcal{S}$, results of modding out shape-preserving transformations (scaling, translations and rotations). The accommodation of ConvNet architectures to properly approximate manifold-valued trajectories on the underlying non-linear space $\mathcal{S}$ is a must. Thus, we make use of geometric steps prior to the encoding-decoding scheme. That is, shape trajectories are first log-mapped to tangent spaces attached to the shape space at a time-varying average trajectory μ, then, obtained vectors are transported to a common tangent space $T_{\mu(0)}(\mathcal{S})$ at the starting point of μ. Without applying any prior temporal alignment (e.g. Dynamic Time Warping) or modeling (e.g. HMM, RNN), the transported trajectories are then fed to a convolutional auto-encoder to build subject-specific latent spaces. The proposed approach was tested on two publicly available datasets. Our approach outperforms existing approaches on CMU gait dataset, while performances on UPCV K2 are comparable to existing approaches. We demonstrate that combining geometric invariance (i.e. Kendall's representation) with our data-driven ConvNet model is suitable to alleviate spatial and temporal variability, respectively.

1. Introduction

During this decade, the automatic estimation of 3D human body skeletal data in video streams, either RGB or RGB-D, have received a particular attention. As a result, real-time and accurate algorithms have been developed and released [25, 24, 32]. The analysis of such data (time-series shape data) allows human behavior understanding as action and activity recognition [7, 29, 30], gait recognition, gender classification, etc. The abundance of this kind of data recently have opened the gate to explore and study their static and dynamic properties and make use of them in application fields as health-care and well-being, gaming, action and activity recognition [30, 7, 29], gait analysis [22] and recognition [5, 6, 13]. Estimated 3D skeletal data have the advantage to handle the camera projection problem often present in the silhouette-based approaches [21]. Extracted Silhouettes are thus distorted by projection on the image plan, which makes their analysis view-dependant. However, with respect to the camera scenes and views, estimated 3D data yield many variations that lead to an unreasonable analysis if used in a raw state. Therefore, one needs to filter out shape-preserving transformations in order to acquire suitable invariance properties [7] required in analyzing 3D skeletons. The elegant and rigorous Kendall's approach [18] ends up with a set of shape orbits invariant to scaling, translation and rotations. However, applying machine learning approaches (such as Deep learning) on such manifold-valued trajectories is not straightforward due to the non-linearity of the underlying space over which the use of euclidean geometric tools and the euclidean metric remains irrelevant. Another challenging problem when comparing shape trajectories in general and gait sequences in particular is the temporal variability. That is, different gait sequences of the same person could be performed at different execution rates. They could also exhibit a temporal shift with each other. This is a classical problem in computer vision, often solved using a temporal warping of the sequences in hand prior to the classification. Dynamic Time Warping and its variants are the most popular solutions (see for example [30]). More elaborated metrics have been proposed as in [7] based on T-SRVF representation and used in [13] for gait recognition. Another alternative will be to use temporal models in the classification schema as Hidden Markov Models (HMM) or Recurrent Neural Networks (RNN). In this work, we demonstrate that a Deep Convolutional Autoencoder (DCAE) trained on gait shape trajectories handles temporal variations. When shape-preserving transformations (scaling, translation and rotation) are filtered out from the static 3D shape repre-

sentation, the temporal variations, as well as other intra-class variations, are accounted using our data-driven model, termed DCAE. To illustrate it experimentally, we consider here the specific problem of gait recognition from 3D skeletal data. Gait sequences are mapped to a shape space $\mathcal{S}$ which builds up time-parameterized trajectories termed $\alpha(t)$, where $t \in [0, 1]$ is the time domain. Coming back to the encoding-decoding problem of shape trajectories $\alpha(t)$, one could view it as solving the following loss function $l_{\mathcal{S}}$,

$$l_{\mathcal{S}}(\alpha) \triangleq \min_{w,b} \left(\int_t d_{\mathcal{S}}(\alpha(t), f(g(\alpha(t))^2 dt \right)$$

that is, one seeks to optimally approximate an arbitrary input $\alpha(t)$ by $\hat{\alpha}(t) = f(g(\alpha(t)))$, grounding on a forward process $g(.)$ by which we aim to project the input trajectory into the low-dimensional latent space and a backward process $f(.)$ by which the input is reconstructed based on its latent representation. The main problem here lies in the fact that the function f uses linear combinations of the inputs which will violate the non-linear structure of $\mathcal{S}$ and obtained reconstructions could step outside the shape manifold $\mathcal{S}$. To overcome this problem, we propose in this work an intrinsic approach grounding on geometric tools explicitly defined on the Kendall shape space of 3D skeletal data [7]. We first compute a sample trajectory $\mu(t)$, similarly to [13]. Second, all skeletal shapes are log-mapped to the tangent bundle $T_{\mu(t)}(\mathcal{S})$ then transported to a common tangent space attached to $\mathcal{S}$ at the initial point $\mu(0)$. Once done, transported trajectories live on the same vector space and conventional encoding-decoding architectures can be safely applied on the reference tangent space $T_{\mu(0)}(\mathcal{S})$. It is possible to project back transported trajectories to the space of interest $\mathcal{S}$ and get approximated reconstructions $\hat{\alpha}(t)$. To summarize, the main contributions of this work are three-folds: **(1)** A novel geometric deep convolutional auto-encoder (DCAE) approach for classifying 3D skeletal shape trajectories. To our knowledge, this is the first attempt to accommodate Deep Learning techniques to the Kendall's shape space. The proposed approach shows robustness to intra-class spatio-temporal variations; **(2)** A comprehensive study of Deep Convolutional Autoencoder (AE) trained on 3D skeletal data in comparison with other three variants including Gentle AE, Deep AE, Convolutional AE for the purpose of 3D gait recognition; **(3)** Extensive experiments on two publicly available datasets, comparative studies with the state-of-the-art and an emphasize on important parameters as the temporal resolution of shape trajectories, the size of the convolution filter and the contribution of shape dynamics.

The rest of the paper is organized as following. In Sec.2, we briefly review existing solutions for 3D Gait recognition as well as Deep Learning approaches applied to 3D skeletal data. Sec.3 introduces our trajectory representation of 3D gait sequences. It reviews geometric properties of the Kendall's shape space as well as key operations. Our geometric encoding-decoding schemes and related 3D gait classifiers are detailed in Sec.4. Experimental results and evaluation discussions are reported in Sec.5. Some conclusions and perspectives are drawn in Sec.7.

2. Related Work

In this section, we review recent 3D gait recognition approaches and Deep Neural Network techniques applied to 3D skeletal data acquired using both MoCap sensors including Kinect-like sensors.

3D Gait Recognition – The majority of existing 3D gait recognition approaches are based on handcrafted features. For instance, based on a 3D volumetric (voxel) gait dataset, authors in [4] have extracted both gait structural and dynamics features by generating an energy map between the data and a structural gait model. Recently, several approaches [1, 2, 6, 26, 20] have exploited 3D skeletal data instead of silhouettes or volumetric data. Presenting a great potential, this kind of data is independent of the illumination conditions, robust to self-occlusions and to pose variations. They are also source of relevant features as the anthropometric measurements (body's height, length of arms and legs, etc.) and kinematic features (the stride length, evolution of some angles, gait patterns, body's velocity, etc.). In literature, these features are either used separately or fused to provide a gait signature used for recognition. For example, Preis et al. [23] extracted thirteen gait features where eleven are static features of the human body (height, length of legs, length of both upper arms, etc.) and two are dynamic features that are step length and the body's speed. In addition to these latter features, Sinha et al. [26] used other gait parameters such as areas of upper and lower body and inter-joint distances. Kwolek et al. [20] also computed inter-joint distances along with bone rotations and person's height. Horizontal distances and vertical ones between joints during one gait cycle were considered in [1]. Statistical tools like the mean and standard deviation were used in [9] and [2] on extracted gait attributes such as joint angles, inter-joint distances or lower limbs angles (hips, knees and ankles) in order to get gait descriptor sets. In [6], as machine learning techniques are well used in 2D gait recognition approaches, Balazia et al. exploit some of these techniques to get better gait classification results. In fact, gait features are learned by maximizing the inter-class separability via a modification of Fisher's Linear Discriminant Analysis with Maximum Margin Criterion (MMC). More recently, Hosni et al. have proposed in [13] to model 3D gait cycles as trajectories on the Kendall's shape space. Grounding on the elastic metric introduced in [7], they extended the well-known functional PCA on shape trajectories mapped onto the tangent bundle attached to the shape manifold on an arbitrary

average gait trajectory. In addition to this latter work, as Kendall's space is a Riemannian manifold, many other researchers actually dealt with skeletal sequences depending on Riemannian modeling assumptions. In fact, as already cited, Ben Amor et al. [7], through introducing the elastic metric, opted to parallel translate trajectories lying on the Kendall space to a reference tangent space attached to the underlying space at a fixed point. One can also cite [30] where authors modeled skeletal sequences as trajectories lying on the Special Euclidean (Lie) group $SE(3)^n$, then they were mapped the tangent space attached to the Lie group at the identity i.e. Lie algebra $\mathfrak{se}(3)^n$ where they were exploited for action recognition.Working on the same manifold, the approach in [3] is based on the TSRVF (Transport Square-Root Velocity Function) representation [28] offering a metric with good properties to overcome the inherent space non-linearity and temporal variability. By overcoming the two latter challenges, they adapted some machine learning methods such as PCA (i.e. Principal Component Analysis) to the underlying space. Recently [16] have collected two datasets called UPCV and UPCV K2. They proposed to capture the deviation of 3D poses with respect to a global model, in addition to intra-sequence pose variability. They then map extracted features in a RKHS (Reproducing Kernel Hilbert Space) of Euclidean and Riemannian features fused using the Handmard product. The identity classification is based on a kernelized version of SRC.

Deep Neural Networks on 3D Motion Data – Other approaches to gait recognition and 3D motions in general are based on deep learning and does not use any handcrafted features. All features are trained via the neural networks that have shown their great power in learning compact and discriminative representations for images and videos, thanks to their ability to perform nonlinear computations. In particularly, convolutional neural networks are now very popular in different computer vision problems related to 3D skeletal data and achieve highest results. Based on [27], their first application to gait recognition was made not long ago in [8] using spatio-temporal cuboids of optical flow as input data. Earlier, while considering a 3D human body sequence as time serie of the joints, many other researchers employed recurrent neural networks (RNNs) with Long-Short Term Memory (LSTM) neurons [[11],[33]]. These architectures presents difficulties to memorize the information of the entire sequence with many timesteps [31]. Holden et al. [12] tried to exploit 3D skeletal data jointly with a Convolutional autoencoder in order to provide a motion manifold permitting synthesis of characters movements. They also used an integration process by stacking a feed-forward network on top of the autoencoder for the sake of producing realistic motion sequences. As a new trend, other works were interested in applying deep learning on non-Euclidean geometric data. For instance, [14] incorporated the Lie group structure into a deep network architecture to learn more appropriate Lie group features for 3D action recognition. The main drawback of this approach is that while it allows to extend CNNs to a non-Euclidean domain, it does not allow applying the same model across different domains, since the convolution coefficients are domain-dependent. Herein, inspired by the work of Holden et al. [12], we are interested in providing a novel method that mixes non-euclidean structure with a deep learning framework. In particular, the main differences of our approach compared to [12] are: (i) they used skeletal data without any manifold assumptions while we modeled them as time-parameterized shape trajectories lying on a non-linear space as described in Sec.3; (ii) they exploited the latent representation for synthesis whereas we use theses features obtained through the encoding-decoding optimization to classify different persons identities (Sec.4).

3. Skeletal Kendall's Shape Trajectories

Among promising approaches for skeletal motion data representation and analysis, one can cite the Kendall's 3D shape trajectory approach proposed in [7] and successfully extended in [29, 13]. The underlying representation build rigorously a shape space where shape-preserving transformations (i.e. scaling, translations and rotations) are filtered out. To reach temporal rate-invariance and shift-invariance, Ben Amor et al. [7] have introduced an elastic metric which accounts for temporal stretching and compression of shape trajectories.

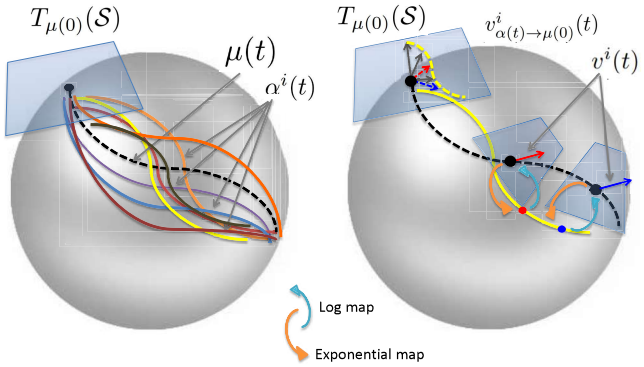

Figure 1. Key geometric steps prior to the Encoding-Decoding using Deep Convolutional Neural Networks. The left panel shows a set of shape trajectories $\alpha^i(t)$ and their mean trajectory $\mu(t)$ (dashed line). In the right panel, $v^i(t)$, the log-mapped versions of $\alpha^i(t)$ to the tangent spaces around the mean trajectory, are transformed into $v^i_{\alpha(t)\to\mu(0)}(t)$, element of $T_{\mu(0)}$ by parallel translation.

This rate-invariant metric allowed Hosni et al. [13] to adapt the well-known fPCA (functional Principal Component Analysis) on shape trajectories and obtain uncorrelated latent variables as new representations. Taking a different direction, [29] have introduced on top of the Kendall's

representation a Sparse Coding and Dictionary Learning (SCDL) approach to effectively represent shape trajectories using time-series with suitable sparse and discriminatory properties. To alleviate the temporal variability, the classification task uses Dynamic Time Warping followed by a one-vs-all SVM classifier or applied a bidirectional LSTM (Bi-LSTM). In this work, in contrast, we propose geometric coding-decoding (AutoEncoder) Neural Networks for effective 3D gait recognition without a prior definition of any elastic metric [7] or temporal warping (e.g. DTW) [30, 29]. Applying AE, or any advanced Deep Convolutional architecture, is not straightforward as trajectory representations lie to a shape space $\mathcal{S}$, a non-linear orbifold (set of orbits of a preshape space $\mathcal{C}$). To overcome this constraint, we exploit geometric properties of the underlying shape space, in particular the parallel translation, the exponential and the logarithm map applications, which will be described in this section. Formally, we shall approximate a sequence of discrete 3D skeletal data $X_{t=0,1/T,2/T..,1}$ with a continuous smooth time-parameterized trajectory $\alpha(t)$, with t in a time domain $[0, 1]$. At any time t, scaling, translation and rotational effects are filtered out to keep only shape-relevant information in $[\alpha(t)]$. After removing the translation then the scaling, trajectories lie to the unit sphere of $\mathbf{R}^{3(n-1)}$, where n denotes the number of landmarks – it is termed a preshape space $\mathcal{C}$. The shape space is defined by $\mathcal{S} = \mathcal{C}/SO(3)$ the quotient space of $\mathcal{C}$ by the rotation group of $\mathbf{R}^3$. For further details on this methodology, we refer the reader to [18, 10] and [7]. Our current goal is to make use of Deep Convolutional Autoencoders for the classification of gait shape trajectories. To this end, we propose to parallel translate log-mapped shape trajectories to a fixed tangent plane attached to $\mathcal{S}$ at the origin of a pre-computed average trajectory (Fig.1) prior to train our Convolutional Autoencoder. In the following, we provide definitions of some useful tools such as tangent plane to $\mathcal{S}$, Exponential and Logarithm maps, and the parallel translation operation.

– Tangent space to $\mathcal{S}$ – The tangent space attached to the pre-shape space at $X \in \mathcal{C}$ is given by $T_X(\mathcal{C}) = \{V \in \mathcal{C} | \text{trace}(V^T X) = 0\}$. Hence, the tangent plane to $\mathcal{S}$ at $[X]$ can be defined as, $T_{[X]}(\mathcal{S}) = \{V \in \mathcal{C} | \text{trace}(V^T X) = 0, \ \text{trace}(V^T X A) = 0\}$, where A is any skew-symmetric matrix of size 3×3. The first condition makes V tangential to the preshape space while the second condition imposes its orthogonality to the rotation orbit. Together, they guarantee V to be tangent to $\mathcal{S}$. For convenience, the tangent space $T_{[X]}(\mathcal{S})$ is identified with $\mathbf{R}^{3n-7}$.

– Exponential map – for $V \in T_{[X]}(\mathcal{S})$, the Exponential function $\exp_{[X]}(.) : T_{[X]}(\mathcal{S}) \to \mathcal{S}$ is defined as,

$$\exp_{[X]}(V) = \left[\cos(\theta) X + \frac{\sin(\theta)}{\theta} V \right]. \qquad (1)$$

where $\theta = \sqrt{V, V} = \sqrt{\text{trace}(V V^T)}$.

– Logarithm map – the inverse of the Exponential map $\exp^{-1}_{[X]}(.) : \mathcal{S} \to T_{[X]}(\mathcal{S})$, is given by V,

$$V = \exp^{-1}_{[X]}([Y]) = \frac{\theta}{\sin(\theta)} (YO^* - \cos(\theta) X) \qquad (2)$$

with $\theta = \cos^{-1}(\text{trace}(X(YO^*)^T))$. Here, O^* is the optimal rotation needed to register Y to X: $O^* = \text{argmin}_{O \in SO(3)} \|X - YO\|^2_F$. O^* is found via a *Procruste Analysis* [10]. This θ is also called *Geodesic distance* as it quantifies the length of a **geodesic (i.e. shortest) path** along the space that connects a source shape $[X] \in \mathcal{S}$ to a target shape $[Y] \in \mathcal{S}$.

– Parallel Translation: Additionally, later on, we will also need to transport tangent vectors from arbitrary points in $\mathcal{S}$ to a reference shape termed $[\alpha(0)] \in \mathcal{S}$. This represents transfer of instantaneous deformation from one source shape to another shape while respecting the geometry of the shape space. The parallel translation of tangent vectors, along a curve, is given by an ODE, as described in [17]. While one can use a numerical implementation with dense time steps to solve this ODE, we follow [7] and use **a coarse approximation** in this paper, to gain speed. For shapes $[X]$ and $[Y]$, and a tangent vector $V \in T_{[X]}(\mathcal{S})$, an approximation of the parallel transport of V to $[Y]$, along a geodesic connecting $[X]$ and $[Y]$, is given by Eq.3,

$$V_{[X] \to [Y]} \approx V - \frac{2VYO^*}{\|X + YO^*\|^2_F} (X + YO^*) \, . \qquad (3)$$

In addition to these operations, one needs to define an average gait trajectory. Following [7], we compute a sample average trajectory $\mu(t)$ as a sequence of cross-sectional mean shape (Karcher mean of a set of shapes $[X_i]$ [10]) at each time t of temporally aligned 3D gait trajectories by minimizing Eq.(4).

$$\hat{\mu}(t) = \underset{[X] \in \mathcal{S}}{\text{argmin}} \sum_{i=1}^{N} d_{\mathcal{S}}([X], [X_i])^2 \qquad (4)$$

4. Encoding of Transported Gait Trajectories

Drawing from the success of principal component analysis as a feature reduction algorithm combined with the breakthroughs of Convolutional neural networks (CNNs) in features extraction and classification, we propose to combine geometric tools related to the kendall's shape space with CNNs' power in particular Autoencoder architechtures.

Based on the definitions cited in the previous section (Section. 3), we present here the key geometric steps prior to the Encoding-Decoding steps, as illustrated in Fig.1. For that, let $\{\alpha^i(t)\}$ be a set of skeletal shape trajectories resampled to be of a fixed temporal length. Again, this operation is inherited from the framework of [7]. Geometric operations followed by the encoding-decoding scheme consist in,

1. Compute a sample average shape trajectory $\mu(t)$ from a set $\{\alpha^i(t)\}^{T_r}$, T_r is the training set;

2. Log-map training T_r and testing T_e trajectories $\alpha^i(t)$ to the tangent spaces $T_{\mu(t)}(\mathcal{S})$, using Eq.(2). Let $v(t)^i \in T_{\mu(t)}(\mathcal{S})$ denote obtained tangent vectors,

3. Parallel translate tangent vectors $v^i(t)$ to $T_{\mu(0)}(\mathcal{S})$, using Eq.3. This led to $v^i_{\alpha(t)\to\mu(0)}(t)$, with suitable vector space properties,

4. Apply variants of AE on $v^i_{\alpha(t)\to\mu(0)}(t)$, elements of $T_{\mu(0)}$ with suitable vector space properties,

5. If needed, one can go back to the original trajectory representation via a reconstruction from latent variables, using the parallel translation (Eq.3) then the exponential map application (Eq.1).

In that, the encoding-decoding problem of trajectories $\alpha(t)$ elements of $\mathcal{S}$ is turned to encoding-decoding of their transported version to $T_{\mu(0)}(\mathcal{S})$ after log-mapping $\alpha(t)$ into corresponding tangent bundle. Once done, reconstructed tangent vectors could be mapped back to the shape space $\mathcal{S}$ to approximate original shape trajectories. As $T_{\mu(0)}(\mathcal{S})$ is a vector space of the same dimension than $\mathcal{S}$, we shall use conventional architectures of Autoencoders that will be described in the following items of this section. We highlight the fact that, prior to the aformentioned steps, no temporal synchronization is applied to the gait trajectories in contrast to [13] that temporally aligned them in order to make inference interpretable while employing functional principal components analysis.

Hereinafter, we provide more details about the proposed encoding-decoding problem notably about steps 4. and 5. Actually, when adapted to our $\alpha(t), t \in [0,1]$, shape trajectories on $\mathcal{S}$, it could be seen as a minimization problem of the following loss function,

$$\begin{aligned} l_{\mathcal{S}}(\alpha, w, b) &\triangleq \min_{W,b} \left(d_T(\alpha, C(\alpha, w, b))^2\right) \\ &\triangleq \min_{W,b} \left(\int_t d_s(\alpha(t), C(\alpha(t), w, b))^2 dt\right) \end{aligned} \tag{5}$$

The above formulation is now well adapted to $\mathcal{S}$ taking into account that the encoding function C reflects the encoding-decoding processes that enable the reconstruction of any input trajectory; in other words, C needs to approximate $f \circ g$. In the Euclidean case, C defines linear combinations of some filter weights matrix W. However, using this operator on shape trajectories will result in outputs that steps out of the shape space $\mathcal{S}$. Here, the distance $d_T = \int_t d_{\mathcal{S}}([X(t)], [Y(t)])$, is well adapted if $C(\alpha(t), w, b)$ is a trajectory on the same manifold $\mathcal{S}$ which is not necessary the case.

Therefore, translating the encoding-decoding problem to Kendall Shape Space can't be directly done. This is mainly due to the non-linear geometry on the underlying space. A seemingly straightforward method to overcome this problem is via an intrinsic approach by applying the above mentioned steps from 1 to 3 described in section.3. Consequently, recasting Eq.(5) according to this embedding, results in

$$\min_{W,b} \|v^i_{\alpha(t)\to\mu(0)}(t) - f(g(v^i_{\alpha(t)\to\mu(0)}(t))\|_2^2 \tag{6}$$

For better clarity, we will use $\tilde{v}^i$ instead of $v^i_{\alpha(t)\to\mu(0)}$ to note the transported version of the trajectory $\alpha(t)$ to $T_{\mu(0)}(\mathcal{S})$. We are interested in training the deep convolutional variant of the autoencoding network. As we jointly optimize the parameters of the encoder g and decoder f over the least-squares reconstruction cost Eq.(6), we need to precise these two operations to train the parameters taking into account that $\tilde{v}^i(t) \in R^{F\times(3N)}$ with F is the temporal resolution and N is the number of 3D skeleton landmarks. Thus, in our experiments $\tilde{v}^i(t)$ are reshaped to stacked 1-D trajectories.

As depicted in Figure. 2, **Deep Convolutional Autoencoder** consists in three alternations between a 1D convolution layer and a 1D maxpooling layer for the encoder and in three alternations between an 1D upsampling layer and a 1D convolution layer for the decoder also with the hyperbolic tangent as an activation function. More specifically, for each alternation, given the convolution operator , max pooling operator Ω, filter weights W_e and biases b_e, the encoding operation is given by the following.

$$h^k = g(\tilde{v}^i(t)) = \Omega(Tanh(\tilde{v}^i(t) * W_e^k + b_e^k)) \tag{7}$$

where Ω allows the filters to express a degree of temporal invariance by reducing the temporal resolution to only representative features [12]. As for the decoding operation, considering the convolution operator *, the upsampling operator Φ, filter weights W_d and biases b_d, it can be written as follows,

$$f(g(\tilde{v}^i(t))) = Tanh(\sum_{k\in\mathbb{H}} \Phi(h^k) * W_d^k + b_d^k) \tag{8}$$

It is as though we are reconstructing the input data by convolving trained filters with latent feature maps.

Accordingly, the autoencoder network, while optimizing Eq. 5 in the training phase, will allow the shape trajectories whether they belong to the training or the testing set to be reconstructed from the latent variables by using the parallel translation (Eq.(3)) then the exponential map application (Eq.1).
Figure. 3 displays an arbitrary trajectory α (from CMU) and its reconstruction $f(g(\alpha))$ using our DCAE. The distance $l_{\mathcal{S}}(\alpha) = 0.084$ on $\mathcal{S}$ is quite low. The trajectories are

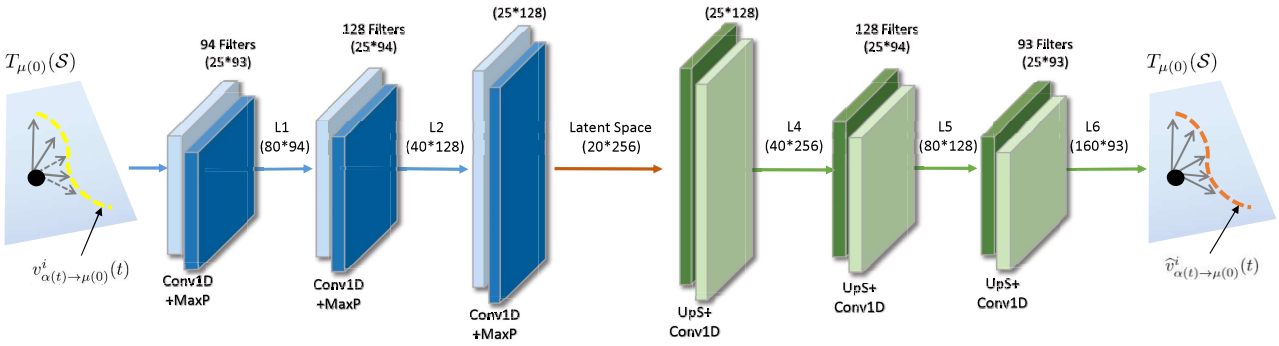

Figure 2. Deep Temporal convolutional autoencoder network architecture. Note that Conv1D stands for 1D convolutional layer that convolve only in the direction of time, MaxP and UpS identifies maxpooling and upsampling layers respectively.The number and size of layers is only for illustrative purposes. Blue layers depict the forward process while the green layers represent the backward process. The first layer for example contains 94 filters of size 25x93. The first dimension of the filter corresponds to a temporal window, while the second dimension corresponds to the number of features/filters on the previous layer. Consequently, after convolving and applying temporal maxpooling, parallel translated tangent vectors $v^i_{\alpha(t)\to\mu(0)}(t)$ of size (160*93), the first dimension is divided by 2 while the second is fixed to the number of filters (80*94).

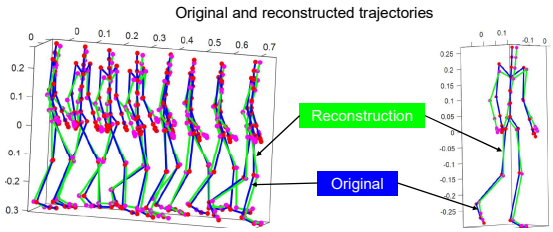

Figure 3. Reconstruction (in green) of an arbitrary sequence from CMU gait dataset (original in blue). Pairwise distance between trajectories is 0.084.

perfectly synchronized which proves the efficiency of convolutional networks in dealing with non aligned shape trajectories. All that being said, we also exploit the forward process of the autoencoder network as a shared network to predict the identities labels. Once training is complete, the filters exhibit strong temporal and inter-joint correspondences. Each filter expresses the movement of several joints over a period of time which corresponds to a natural, independent components of gait motion. In fact, grounding on the output $\mathbb{H}$ that presents optimal latent codes capturing kind of meaningful representation of our data, we added a softmax layer which allows us to deal with the classification task with sufficient predictive power.

5. Experimental Results

In this section, we discuss several evaluation results of our approach w.r.t. existing approaches. We consider two publicly available 3D gait datasets – (1) the gait subset of the Mocap CMU dataset[1]; and (2) the UPCV K2 collected by the authors of [16]. On the first dataset, we report the results of our Transported DCAE and three other variants of the AE (i.e. Gentle AE, DAE and CAE). We study the impact of the temporal resolution of gait trajectories and the size of the convolutional filters. On UPCV K2, we report results of our DCAE. Finally, we report an ablation study which demonstrates the superiority of our Transported DCAE compared to LSTM/Bi-LSTM and DTW and shows the importance of the Kendall's shape representation.

5.1. Evaluation on CMU MoCap Gait Dataset

The CMU MoCap dataset contains different human actions and daily activities such as walking, running, playing tennis, etc. It includes $3,843$ gait cycles of 54 subjects. They were extracted and released[2] by Balazia et al. [6]. We have followed the homogeneous experimental protocol described in [6] to train a shared discriminant network with variants of AE as described in Section. 4. In details, a 3-fold cross validation is performed by splitting extracted gait cycles to three folds: one training set and two evaluation sets. Each set contains disjoint instances from all the subject classes. Thus, the data of training set are only used to compute the average sample trajectories then the evaluation sets are divided into one fold as a test set and nine others as a gallery set based on nested 10-fold cross validation. We trained our AE networks for 50 epochs and make use of the adaptive gradient descent algorithm, i.e Adam. While evaluating the shared network in terms of Correct Classification Rate (CCR), our approach outperformed the state-of-art as reported in Table. 1. Without any temporal alignment (e.g. DTW) or modeling (using HMM or RNN) of gait trajectories, an improvement of more than 3% is reported which represents more that 76 additional sequences

[1] http://mocap.cs.cmu.edu

[2] https://gait.fi.muni.cz/#framework

Table 1. Performances on CMU MoCap gait dataset and comparison with respect to state-of-the-art.

Method	Year	# of features	CCR
Preis et al. [23]	2012	13	0.1300
Sinha et al. [26]	2012	45	0.7666
Ahmed et al. [1]	2014	24	0.7134
Dikovski et al. [9]	2014	71	0.8926
Kwolek et al. [20]	2014	660	0.9099
Andersson et al. [2]	2015	80	0.7787
Balazia et al. [6] (PCA+LDA)	2016	54	0.8314
Balazia et al. [6] (MMC)	2016	53	0.9102
Hosni et al. [13] (fPCA+SVM)	2018	85	0.9223
Ours (AE)	–	200	0.9412
Ours (DAE)	–	128	0.9399
Ours (CAE)	–	160×94	0.9354
Ours (DCAE)	–	**20×256**	**0.9597**

correctly classified by our approach. This demonstrates its ability to compensate for temporal shift and rate variability which characterize any gait cycles especially compared to [13] that opted for a registration step when applying functional PCA on shape trajectories.

– Impact of the temporal resolution – We have undertaken some experiments to study the impact of varying the temporal resolution on the correct classification rate when using the different variants of encoding. The graph illustrated in the right panel of Figure. 4 shows that up to 80 frames per sequence there is a considerate increase in the CCR for all the variants to reach a CCR around the 90% then it varies slightly to attain a CCR value of 95.97% for the deep convolutional autoencoder with 160-frames sequences using filters of width 25. We specify that, for these experiments, the convolution filter size is proportional to the temporal resolution, in that 25 temporal width for 160-frames trajectories, 12 temporal width for 80-frames trajectories and so on. These results prove that we need a sufficient number of frames per sequence in order to capture fluent gait patterns with stance and swing phases' details. It could be said that a natural time evolution of walking is needed to be able to catch the appropriate dynamics properties in the latent representation thus improving the classification task performance.

– Which contribution of the shape dynamics? – While modeling the gait sequences as shape trajectories, it is interesting to study the impact of the dynamics in our approach. As a matter of fact, given 160-frames trajectories, we consider only a portion (10%, 25%, 50%, 75% and 100%) of the data either to train the models or to evaluate them. The obtained results are presented in left panel of Figure. 4. For 10% of the trajectories, the CCR values are the lowest for each AE variant. Considering 25% and 50% and 100% of gait trajectories, CCR keep increasing to reach about 95.97%. Based on these observations, one can highlight the importance of the gait dynamics to predict the subject's identity considering the fact that the latent space features are learned by optimizing the autoencoding objective (Eq. (6)). Moreover, knowing that stance phase represents 60% of a stride, it provides most of the discriminating information but the swing phase also contributes to achieve higher performance. The more dynamics information we get, the better is the performance of gait recognition task.

– Impact of the convolutional filter size – We have carried out experiments when varying the convolutional filter size. We have studied in particular the DCAE (Deep Convolutional Autoencoder). We have fixed the temporal resolution to 160 frames per cycle. Reported results are shown in Table. 2. One can observe that the results are similar for 5 and 15 filter sizes, increase with convolutional filters of size **25** then decreases slightly for sizes 30, 40 and 45. Taking into account that a gait cycle is a union of a stance phase and a swing phase and the sequences are not synchronized, we assume that discriminant information is captured during the stance phase. So, if we set the filter width is too low, it is as if we are training the model per frame and if it is the opposite, it is learning a one block motion.

Table 2. Impact of the convolution filter size in the convolutional layers in the case of DCAE.

Filter size	5	15	**25**	30	40	45
CCR (%)	93.68	93.85	**95.97**	94.19	94.79	94.67

5.2. Evaluation on UPCV K2 Dataset

Unlike the CMU MoCap dataset, the UPCV K2 [16] was collected using a Kinect V2 sensor. The sensor was placed at 30 degrees relative to the walking line. The dataset consists of 300 sequences of 30 subjects (17 males and 13 females). Following the experimental settings reported in [16], we have split the dataset into 20% of samples per person for test and the rest 80% for train. Then, we conducted this experiment for 20 iterations using optimal parameters obtained from previous experiments reported in Sec. 5.1. Obtained results are disclosed in Table. 3 with respect to state-of-the-art approaches. They show respectful performance however it still falls behind [16] and [13] results. We emphasize that we were able to reach an accuracy of 92.41% by adding a batch-normalization layer to the proposed architecture based on [15]. This have maybe alleviate the over-fitting effect caused by the reduced size of this database: the model is generalizing better without learning too well the details and the noise from training data.

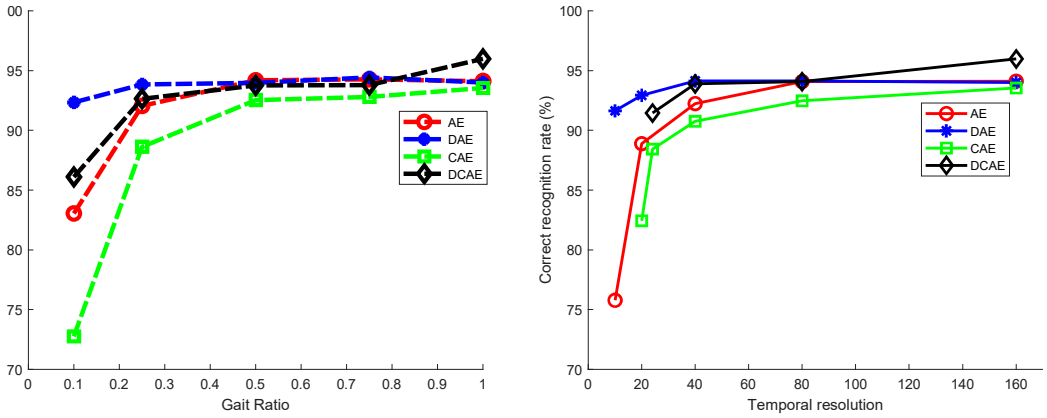

Figure 4. Left panel: impact of the dynamics (gait ratio) on the recognition accuracy; Right panel: impact of the temporal resolution on the recognition accuracy for the four AE variants.

Table 3. Evaluation on UPCV K2 dataset.

Method	CCR
Preis et al. [23] (from [16])	0.4563
Kumar et al. [19] (from [16])	0.9017
Kastaniotis et al. [16] (RK only)	0.9333
Kastaniotis et al. [16] (EK only)	0.9617
Kastaniotis et al. [16] (EK+RK)	0.9705
Hosni al. [13] (fPCA+SVM)	0.8842
Ours (DCAE, filter size=25)	0.8500
Ours (DCAE with BN)	**0.9241**

6. Ablative study

In addition to the above cited investigations, we have conducted experiments based on DTW+NN (Nearest Neighbor), LSTM (Long short-term memory) and Bi-directional LSTM. As shown in Table. 4, the results reveal the limitations of Recurrent models (LSTM/Bi-LSTM) since all gait sequences are similar, differently to human actions or activities. These results also state the interest of our DCAE compared to a temporal alignment performed using the well-known DTW algorithm.

Table 4. Comparison of our approach with baseline algorithms (LSTM, DTW, and without geometric normalisation).

Approach - CCR (%)	CMU gait	UPCV K2
LSTM	36.19	18.25
Bidirectional LSTM	46.45	30.83
DTW+Nearest Neighbor	76.10	71.66
No shape normalisation	92.49	68.74
without Parallel Transport	94.47	84.00
Ours (DCAE)	**95.97**	**85.00**

From another perspective, the second part of Table. 4 highlights the merit of our geometric pipeline – shape normalization (Kendall's representation), Riemannian log-map and Parallel Translation – and the geometric extension of DCAE. When applied on translated log-mapped data, our DCAE achieves higher performances compared to DCAE applied to skeletal data (i.e. without shape normalization). An improvement of 3% on CMU (resp. 9% on UPCV K2) is achieved by the latter one. Besides, we note that the proposed approach performs less better when only applied to log-mapped data on the tangent bundle. To sum up, the intra-class variability (both spatial and temporal) is handled by a two-step strategy. The first one is based on a geometric normalization to filter out shape-preserving transformations (scaling, translations and rotations) which ends up with the Kendall's shape representation. The second strategy, in contrast, allows robustness to temporal variations based on data-driven invariance.

7. Conclusion and Future work

In this work, we have introduced a novel geometric deep convolutional encoding-decoding networks on the Kendall's Shape Space for the purpose of 3D gait recognition. We opted for an intrinsic approach to overcome the non-linearity constraint of geometrically normalized data lying on the underlying space. To overcome the non-linearity of the shape space, transporting the original trajectories to a common tangent space was performed. Experimental results on two publicly available datasets show the competitiveness of the proposed approach compared to existing studies. When rigid transformations of human shapes are filtered mathematically (Kendall's representation), temporal variations are handled thanks to the temporal ConvNet architecture.

Acknowledgment

This work received the financial support of the PHC Utique program for the DEFI project #16G1403.

References

[1] Mohammed Ahmed, Naseer Al-Jawad, and Azhin T Sabir. Gait recognition based on kinect sensor. In *Real-Time Image and Video Processing 2014*, volume 9139, page 91390B. International Society for Optics and Photonics, 2014. 2, 7

[2] Virginia O. Andersson and Ricardo M. Araujo. Person identification using anthropometric and gait data from kinect sensor. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 425–431. AAAI Press, 2015. 2, 7

[3] Rushil Anirudh, Pavan Turaga, Jingyong Su, and Anuj Srivastava. Elastic functional coding of human actions: From vector-fields to latent variables. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3147–3155, 2015. 3

[4] G. Ariyanto and M. S. Nixon. Model-based 3d gait biometrics. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7, 2011. 2

[5] Gunawan Ariyanto and Mark S. Nixon. Marionette mass-spring model for 3d gait biometrics. In *5th IAPR International Conference on Biometrics, ICB 2012, New Delhi, India, March 29 - April 1, 2012*, pages 354–359, 2012. 1

[6] Michal Balazia and Petr Sojka. Learning robust features for gait recognition by maximum margin criterion. pages 901–906. IEEE, 2016. 1, 2, 6, 7

[7] Boulbaba Ben Amor, Jingyong Su, and Anuj Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):1–13, 2016. 1, 2, 3, 4

[8] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil Mata, and Nicolas Pérez de la Blanca. Automatic learning of gait signatures for people identification. In *IWANN*, 2017. 3

[9] Bojan Dikovski, Gjorgji Madjarov, and Dejan Gjorgjevikj. Evaluation of different feature sets for gait recognition using skeletal data from kinect. *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th Inter*, May 2014. 2, 7

[10] Ian L. Dryden and Kanti V. Mardia. *Statistical shape analysis*. Wiley, 1998. 4

[11] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 06 2015. 3

[12] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):138, 2016. 3, 5

[13] Nadia Hosni, Hassen Drira, Faten Chaieb, and Boulbaba Ben Amor. 3d gait recognition based on functional pca on kendall's shape space. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2130–2135. IEEE, 2018. 1, 2, 3, 5, 7, 8

[14] Zhiwu Huang, Chengde Wan, Thomas Probst, and Luc Van Gool. Deep learning on lie groups for skeleton-based action recognition. *CoRR*, abs/1612.05877, 2016. 3

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 7

[16] Dimitris Kastaniotis, Ilias Theodorakopoulos, George Economou, and Spiros Fotopoulos. Gait based recognition via fusing information from euclidean and riemannian manifolds. *Pattern Recogn. Lett.*, 84(C):245–251, Dec. 2016. 3, 6, 7, 8

[17] D.G. Kendall, D. Barden, T.K. Carne, and H. Le. *Shape and Shape Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. 4

[18] David G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984. 1, 4

[19] M. S. Naresh Kumar and R. Venkatesh Babu. Human gait recognition using depth camera: A covariance based approach. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '12, pages 20:1–20:6, New York, NY, USA, 2012. ACM. 8

[20] Bogdan Kwolek, Tomasz Krzeszowski, Agnieszka Michalczuk, and Henryk Josinski. *3D Gait Recognition Using Spatio-Temporal Motion Descriptors", year=2014, publisher=Springer International Publishing, address=Cham, pages=595–604,*. 2, 7

[21] Mark S Nixon, Tieniu Tan, and Rama Chellappa. *Human identification based on gait*. Springer, 2006. 1

[22] Alexandra Pfister, Alexandre M West, Shaw Bronner, and Jack Adam Noah. Comparative abilities of microsoft kinect and vicon 3d motion capture for gait analysis. *Journal of medical engineering & technology*, 38(5):274–280, 2014. 1

[23] Johannes Preis, Moritz Kessel, Martin Werner, and Claudia Linnhoff-Popien. Gait recognition with kinect. 01 2012. 2, 7, 8

[24] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3116–3124, USA, 2016. Curran Associates Inc. 1

[25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1297–1304, 2011. 1

[26] Aniruddha Sinha, Kingshuk Chakravarty, and Brojeshwar Bhowmick. Person identification using skeleton information from kinect. In *In The Sixth International Conference on Advances in Computer-Human Interactions*, pages 101–108, 01 2013. 2, 7

[27] A. Sokolova and A. Konushin. Gait recognition based on convolutional neural networks. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W4:207–212, 2017. 3

[28] Jingyong Su, Sebastian Kurtek, Eric Klassen, and Anuj Srivastava. Statistical analysis of trajectories on riemannian manifolds: Bird migration, hurricane tracking, and video surveillance. *Annals of Applied Statistics*, 2013. 3

[29] Amor Ben Tanfous, Hassen Drira, and Boulbaba Ben Amor. Coding kendall's shape trajectories for 3d action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2840–2849, 2018. 1, 3, 4

[30] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, June 2014. 1, 3, 4

[31] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014. 3

[32] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018. 1

[33] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. 2016. 3