This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3291663

Geometric Deep Neural Network Using Rigid and Non-rigid Transformations for Landmark-based Human Behavior Analysis

Rasha Friji, Faten Chaieb, Hassen Drira, *Senior Member, IEEE,* and Sebastian Kurtek, *Senior Member, IEEE*

Abstract—Deep learning architectures, albeit successful in most computer vision tasks, were designed for data with an underlying Euclidean structure, which is not usually fulfilled since pre-processed data may lie on a non-linear space. In this paper, we propose a geometric deep learning approach using rigid and non-rigid transformations, named KShapenet, for 2D and 3D landmark-based human motion analysis. Landmark configuration sequences are first modeled as trajectories on Kendall's shape space and then mapped to a linear tangent space. The resulting structured data are then input to a deep learning architecture, which includes a layer that optimizes over rigid and non-rigid transformations of landmark configurations, followed by a CNN-LSTM network. We apply KShapenet to 3D human landmark sequences for action and gait recognition, and 2D facial landmark sequences for expression recognition, and demonstrate the competitiveness of the proposed approach with respect to state-of-the-art.

Index Terms—Geometric deep learning, human behavior analysis, Kendall shape space, transformation layer.

1 Introduction

HUMAN behavior analysis via diverse data types has emerged as an active research issue in computer vision due to 1) the wide spectrum of not yet fully explored application domains, e.g., human-computer interaction, intelligent surveillance security, virtual reality, etc., and 2) the development of advanced sensors such as Intel RealSense, Asus Xtion and the Microsoft Kinect [1], which yield various data modalities, e.g., RGB and depth image sequences, and videos. Conventionally, these modalities have been utilized solely [2], [3], or merged (e.g., RGB + optical flow), for the action recognition [4], [5], gait recognition and facial expression recognition tasks using multiple classification techniques, and have resulted in excellent results. With the development of human pose estimation [6], [7] and facial landmark detection [8], [9] algorithms, the problem of human landmark (or key-point) localization was solved, and reliable acquisition of accurate 2D/3D landmark data became possible. In comparison with former modalities, landmark data, a topological representation of the human body or face using key-points, appears to be less computationally expensive, and more robust in front of intricate backgrounds and with respect to variable conditions including viewpoints, scales and motion speeds [10]. An efficient way to analyze datasets composed of 2D/3D landmark observations is to consider their shapes independently of undesirable transformations; the resulting representation space of landmark data is non-linear.

Accordingly, we represent 2D/3D landmarks in the Kendall shape space [11] that defines shape as the geometric information that remains after location, scaling and rotational effects are filtered out. A sequence of landmarks is then modeled as a trajectory on this space. Thus, to analyze and classify such data, it is more suitable to consider the geometry of the underlying space. This remains a challenging problem since most commonly used techniques were designed for linear data. Deep learning architectures, despite their efficiency in many computer vision applications, usually ignore the geometry of the underlying data space. Therefore, geometric deep learning architectures have been introduced to remedy this issue.

To the best of our knowledge, the main previous geometric deep learning approaches on manifolds were designed on feature spaces (e.g., space of symmetric positive definite (SPD) matrices, Grassmann manifold, Lie groups [12], [13]) or on the 3D human body manifold [14], [15]. The literature that considers this problem on shape spaces is scarce. An extension of a conventional deep architecture on Kendall's pre-shape space has been recently proposed in [16], and an auto encoder-decoder has been extended to a shape space for gait analysis in [17].

In this work, we extend the KShapeNet geometric deep learning approach on Kendall's shape space [18] to 2D or 3D landmark sequences for human motion analysis. In [18], we considered the problem of 3D skeleton-based action recognition. Here, we further adapt this framework to the problems of gait (3D landmark sequences) and facial expression (2D landmark

- *R.Friji is member of CRISTAL laboratory - ENSI, Université Manouba Campus, Manouba, Tunisie and member of Talan Tunisia, 10 Rue de l'énergie solaire Impasse N°1 Charguia 1, Tunis 2035, Tunisie E-mail: racha.friji@talan.com*
- *F.Chaieb is member of Efrei Research Lab, Paris Panthéon-Assas University and member of CRISTAL laboratory - ENSI, Université Manouba Campus, Manouba, Tunisie. E-mail: faten.chakchouk@efrei.fr*
- *H.Drira is member of ICube UMR 7357, CNRS, Université de Strasbourg, France. E-mail: hdrira@unistra.fr*
- *S.Kurtek is member of the Department of Statistics, The Ohio State University, Columbus, OH, USA. E-mail: kurtek.1@stat.osu.edu*

sequences) recognition. The generality of KShapeNet allows us to address various recognition tasks based on landmark sequences.

Landmark sequences (representing the human body or face) are first modeled as trajectories on Kendall's shape space by filtering out scale and rigid transformations. Then, the sequences are mapped to a linear tangent space and the resulting structured data are input to a deep learning architecture. The latter includes a novel layer that learns the best rigid or non-rigid transformation to be applied to the landmark configurations to accurately recognize the motion. In light of this, our main contributions are as follows.

1) We define a novel deep architecture on Kendall's shape space for landmark-based human motion analysis. In particular, we extend the framework introduced in [18] for action recognition based on 3D landmark sequences to gait and facial expression recognition tasks based on 3D and 2D landmark sequences, respectively.
2) The proposed deep network includes a novel transformation layer that optimizes over rigid and non-rigid transformations of landmark configurations, which increases recognition accuracy for human motion analysis.
3) The proposed architecture is applied to 3D landmark-based motion analysis, namely action recognition and gait recognition, as well as 2D landmark-based motion analysis, namely facial expression recognition. We report state-of-the-art results on five large scale publicly available datasets: NTU-RGB+D and NTU-RGB+D120 for 3D action recognition, CMU Mocap dataset for 3D gait recognition, and CK+ and Oulu-CASIA datasets for 2D facial expression recognition.

The rest of the paper is organized as follows. In Section 2, we briefly review existing research on action recognition, gait recognition, facial expression recognition and geometric deep learning. Section 3 describes geometric modeling of landmark configuration trajectories on Kendall's shape space. In Section 4, we introduce the proposed geometric deep architecture, KShapeNet. Experimental settings, results and discussions for human action recognition, gait recognition and facial expression recognition are provided in Sections 5.1, 5.2 and 5.3, respectively. Finally, Section 6 concludes the paper and summarizes directions for future work.

2 Related work

2.1 Human action recognition

Presently, deep learning methods for human action recognition are preferred over traditional skeleton-based ones, which tend to focus on extracting hand crafted features [19], [20]. The former methods can be categorized into three major sets: methods based on a Recurrent Neural Network (RNN) [21], methods based on a Convolutional Neural Network (CNN) [22], and methods based on a Graph Convolutional Network (GCN) [23].

Since RNNs are convenient for time series data processing, RNN-based methods consider skeleton sequences as time series of coordinates of the joints (landmarks). For the purpose of improving the capability of learning the temporal context of skeleton landmark sequences, Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been introduced as efficient alternatives for skeleton landmark-based action recognition. Zhu et al. [24] used an LSTM network and characterized joints through the co-occurrence between actions. In [25], geometric joint features were applied to a multi-layered LSTM network instead of directly passing in the joint positions. The pitfall of some of these methods [26], [27] is their weak ability in spatial modeling, resulting in non-competitive results. A novel two-stream RNN architecture was recently proposed by Hong and Liang [28]. This architecture models both the temporal dynamics and spatial configurations of skeleton landmark data by applying an exchange of the skeleton axes at data level pre-processing. Relatedly, Jun and Amir [29] focused on extracting the hidden relationship between the two domains, spatial and temporal, using a traversal approach on a given skeleton landmark sequence. Unlike the general method where joints are arranged in a simple chain ignoring kinetic dependency relations between adjacent joints, this tree structure-based traversal does not add false connections between body joints when their relation is not strong enough. Many RNN-based methods in this context suffer from issues related to the gradient exploding or vanishing over layers. Some new RNN architectures [30], [31] were proposed to address this particular limitation.

CNN models have excellent capability to extract high level information and semantic cues. Multiple works [32], [33], [34] have exploited CNN models for action recognition by encoding the skeleton joints as images or pseudo-images prior to feeding them to the network. In [34], Zhang et al. map a skeleton landmark sequence to an image, referred to as the skeleton map, to facilitate spatio-temporal dynamics modeling via the ConvNet. The challenge with CNN-based methods is the extraction and utilization of spatial as well as temporal information from 3D skeleton landmark sequences. Several other problems hinder these techniques, including model size and speed [35], occlusions, CNN architecture definition [36], and viewpoint variation [34]. Skeleton landmark-based action recognition using CNNs, that overcomes these challenges, thus remains an open research question.

Recently, the GCN has been adapted to action recognition. This network represents human 3D skeleton data as a graph (landmarks and connections between them). There are two main types of graph related neural networks: the graph recurrent neural network, and the graph convolutional neural network [37], [38].

2.2 Gait recognition

Gait is an appealing biometric modality. Gait recognition aims to identify individuals based on the way they walk. Thus, several approaches to this problem have been proposed in different application areas including user identification [39], [40], sport science [41], and healthcare [42], [43]. From the perspective of body representation taxonomy, these approaches can be organized into two categories: silhouettes or skeletons. The second category exploits motion capture technology to estimate 3D positions of anatomical landmarks. As opposed to silhouette-based methods, skeleton-based gait recognition methods are generally more robust against viewpoint and appearance changes [44]. 3D skeletal data allow the extraction of static and dynamic anthropometric and gait features, e.g., stride length, speed, distances and angles between joints, from the body joints and their connections [45]. For clustering, the authors in [46] select the mean, standard deviation and maximum value of three angles for each of the left and right legs, hips, knees and ankles. Ding et al. [47] extract Horizontal Distance Features (HDFs) and Vertical Distance Features (VDFs), based on distances between ankles, knees, hands and shoulders. Ortiz et al. [48] compute the mean and standard deviation in the signals of lower joint (hips, knees and ankles) angles. The authors in [49] use geometric features, such as body-derived parameters, joint angles

and inter-joint distances, along with various statistics, to construct seven different feature sets. Preis et al. [50] define 13 biometric attributes among which 11 are static body parameters and two are dynamic parameters (step length and speed). In [51], Sinha et al. combine features introduced in [46] and [50] with a set of other gait features, namely areas of upper and lower body and inter-joint distances. The authors in [52] consider skeleton sequences as trajectories parameterized by time and use Functional Principal Component Analysis (FPCA) to create uncorrelated variables for identity classification from gait data.

2.3 Facial expression recognition

Facial expression recognition refers to the identification of basic emotions, e.g, fear, sadness, disgust, etc., from videos of faces. The availability of reliable facial landmark detectors [8], [9], and the potential applications, prompted development of many ad-hoc approaches [53] for this task. Recent works increasingly focus on addressing facial expression recognition with the use of deep neural networks, mainly CNNs [54] and RNNs [55]. Among these, many approaches exploit geometric features, i.e, summaries extracted from the locations of salient facial landmarks. Wang et al. [56] proposed a unified probabilistic framework built on top of an Interval Temporal Bayesian Network (ITBN) allowing the representation of spatial dependence among the facial landmarks' movements. In [57], the authors proposed two deep networks: the Deep Temporal Appearance Network (DTAN), used to extract temporal appearance features, and the Deep Temporal Geometry Network (DTGN), which captures geometric information about the motion of facial landmarks. These two models were integrated in a network called the Deep Temporal Appearance-Geometry Network (DTAGN). Aiming to capture subtle facial motions, Jain et al. [58] modeled temporal dynamics of face shapes and proposed a new recognition approach using discriminative Latent-Dynamic Conditional Random Fields (LDCRFs). As another geometric approach [59], 2D facial landmarks were represented as time parameterized trajectories via a mapping into the Riemannian manifold of positive semi-definite matrices. A geometry aware dissimilarity measure, provided by temporal alignment between trajectories, was then used to train a pairwise proximity function SVM (ppfSVM) classifier.

Apropos of deep learning methods, Kim et al. [60] proposed to first learn the spatial feature representation of the micro-expression using a CNN. These features of all input frames were then encoded using the LSTM network. In [61], a two-dimensional landmark feature for effectively recognizing facial micro-expression was proposed. This landmark feature, defined by expressing relative distances between facial landmarks, was used as an input image to a CNN-LSTM-based classifier. Similarly, focusing on micro-expressions, Tanfous et al. [62] proposed to encode 2D facial trajectories using Riemannian extrinsic sparse coding and dictionary learning (SCDL). These sparse time series were then classified using a Bi-LSTM network.

2.4 Geometric deep learning

Compared to previous techniques, geometric deep learning is a nascent research area. As mentioned earlier, it studies the extension of existing deep learning frameworks and algorithms to effectively process graph and manifold data. Some manifold-based techniques have proven their success in 3D human action recognition due to view invariance of the manifold-based representation of skeletal data. As examples, we cite the projection on Riemanian manifold [16], shape silhouettes in Kendall's shape space [63], and linear dynamical systems on the Grassmann manifold [64]. Geometric deep learning approaches can be categorized into two main classes: approaches on manifolds and approaches on graphs. This paper is related to deep approaches on manifolds, and thus, we give a quick review of the state-of-the-art in this category.

Manifold-based geometric deep learning approaches extend deep architectures to Riemannian manifolds, interpreted either as feature spaces [12], [65], [13] or the human body shape, i.e., the human body is viewed as a manifold [14], [15]. Huang et al. proposed several networks on non-linear manifolds. In [12], they introduced the first network architecture to perform deep learning on the Grassmann manifold. They presented competitive results on three datasets of emotion recognition, action recognition and face verification. Along similar lines, an architecture on the manifold of SPD matrices was proposed in [65], and similar experimental evaluation proved the effectiveness of this approach. Recently, the same authors proposed an architecture on Lie groups with application to skeleton-based action recognition [13]. These approaches investigated the non-linearity of various feature spaces, but did not consider shape spaces. Limited efforts have recently been made to design deep architectures on some shape-preshape spaces. Friji et al. [16] proposed a deep architecture on the sphere for modeling unit-norm skeletons with application to action recognition. Along similar lines, Hosni et al. [17] extended the auto-encoder to a shape space with application to gait recognition.

3 Modeling of Shape Space Trajectories

Shape spaces are abstract representation spaces on which each point is a specific shape and the distance between two such points captures the magnitude of shape discrepancies between the respective shapes. Kendall's shape space theory [11] defines shape as the property of an object that remains after variations due to translation, scale and rotation are factored out. Let $\{X^t\}_{t\in\{1,\cdots,p\}}$ denote a sequence of p sets of n k-dimensional landmarks corresponding to a human motion. First, we perform data interpolation via cubic splines, to have the same number of frames for each sequence, rather than the commonly used zero-padding technique. Next, we briefly describe the mathematical framework behind Kendall's definition of shape, and the associated shape space.

Let $X = X^t \in \mathbb{R}^{k\times n}$ denote a set of n landmarks in $\mathbb{R}^k$, $k = 2, 3$ at time t. The shape of the landmark configuration X, as proposed by Kendall [11], is extracted by filtering out all shape-preserving transformations: translation, rotation, and global scaling. Translation and scale variabilities can be removed from the representation space via normalization as follows. Let H denote the $(n-1)\times n$ sub-matrix of a Helmert matrix, as detailed in [66], where the first row is removed. In order to center a 2D/3D landmark configuration X, we pre-multiply it by H, $HX \in \mathbb{R}^{(n-1)\times k}$; then, HX contains the centered Euclidean coordinates of X. Let $C_0 = \{HX \in \mathbb{R}^{(n-1)\times k} | X \in \mathbb{R}^{n\times k}\}$, which can be identified with $\mathbb{R}^{k(n-1)}$, the $k(n-1)$ dimensional vector space. Using the standard Euclidean inner product (norm) on C_0, we scale all centered landmark configurations to have unit norm. As a result, we define the pre-shape space as $C = \{HX \in C_0 | \|HX\|^2 = (HX)^T(HX) = 1\}$; due to the unit norm constraint, C is a $(kn-(k+1))$-dimensional unit sphere in $\mathbb{R}^{k(n-1)}$. Henceforth, we will refer to an element of C as $\tilde{X}$, i.e., a centered and unit norm landmark

configuration. The tangent space at any pre-shape $\tilde{X}$ is given by $T_{\tilde{X}}(C) = \{V \in \mathbb{R}^{kn-(k+1)} | \langle V, \tilde{X} \rangle = V^T \tilde{X} = 0\}$; this tangent space captures all possible perturbations of the pre-shape $\tilde{X}$.

A sequence of p normalized k-dimensional landmark configurations, $\{\tilde{X}^t\}_{t \in \{1,\cdots,p\}}$, is considered as a trajectory on C. In subsequent analysis, our representation of landmark sequences further passes to a tangent space. Thus, it is useful to define three Riemannian geometric tools [67], [68] that allow one to map points 1) from the pre-shape space to a tangent space, 2) from a tangent space to the pre-shape space, and 3) between different tangent spaces. Task 1) can be achieved via the logarithmic map, $log_{\tilde{X}} : C \rightarrow T_{\tilde{X}}(C)$, defined as (for $\tilde{X},\ \tilde{Y} \in C$):

$$log_{\tilde{X}}(\tilde{Y}) = \frac{\theta}{\sin(\theta)}(\tilde{Y} - \cos(\theta)\tilde{X}), \tag{1}$$

where $\theta = \cos^{-1}\left(\langle \tilde{X}, \tilde{Y} \rangle\right)$ is the arc-length distance between $\tilde{X}$ and $\tilde{Y}$ on C. Task 2) is carried out via the exponential map, $exp_{\tilde{X}} : T_{\tilde{X}}(C) \rightarrow C$, defined as (for $\tilde{X} \in C$ and $V \in T_{\tilde{X}}(C)$):

$$\tilde{Y} = \cos(\|V\|)\tilde{X} + \sin(\|V\|)\frac{V}{\|V\|}, \tag{2}$$

where $\|V\| = \sqrt{V^T V}$ as before. Finally, for task 3), we use parallel transport, which in short defines an isometric mapping between tangent spaces. The parallel transport along a geodesic path from $\tilde{X}$ to $\tilde{Y}$ on C, $PT_{\tilde{X} \rightarrow \tilde{Y}} : T_{\tilde{X}}(C) \rightarrow T_{\tilde{Y}}(C)$ is defined as (for $\tilde{X},\ \tilde{Y} \in C$ and $U \in T_{\tilde{X}}(C)$):

$$PT_{\tilde{X} \rightarrow \tilde{Y}}(U) = U - \frac{\langle log_{\tilde{X}}(\tilde{Y}), U \rangle}{\theta}\left(log_{\tilde{Y}}(\tilde{X}) + log_{\tilde{X}}(\tilde{Y})\right), \tag{3}$$

where $\langle \cdot, \cdot \rangle$ and θ are the standard Euclidean inner product and the distance between $\tilde{X}$ and $\tilde{Y}$ on C, respectively, as before.

While translation and scale can be dealt with through normalization, rotation variability in Kendall's framework is removed algebraically using the notion of equivalence classes. The rotation group in $\mathbb{R}^k$ is given by $SO(k) = \{O \in \mathbb{R}^{k \times k} | O^T O = I,\ det(O) = 1\}$. For $O \in SO(k)$ and $\tilde{X} \in C$, the action of the rotation group is given by matrix multiplication, i.e., $O\tilde{X}$ is a rotation of $\tilde{X}$. Let $[\tilde{X}] = \{O\tilde{X} | O \in SO(k),\ \tilde{X} \in C\}$ denote an equivalence class of a pre-shape $\tilde{X}$. Then, Kendall's shape space is the quotient space $C/SO(k)$. Rotation variability is removed in a pairwise manner (or with respect to a given template), by optimally aligning two configurations $\tilde{X}$ and $\tilde{Y}$ via Procrustes analysis [66]; we omit the details of this process here for brevity. After optimal rotation, one can use the same Riemannian geometric tools as on the pre-shape space C, e.g., Equations 1-3, to model shapes of landmark configurations.

4 Shape space deep architecture

An overview of the proposed deep learning architecture on Kendall's shape space for human landmark-based motion analysis is given in Fig. 1. 2D/3D landmark configuration sequences are first modeled as trajectories on C, after which each landmark configuration $\tilde{X}$ is mapped to a common tangent space $T_{\tilde{X}_0}(C)$ at a reference shape $\tilde{X}_0$. The reference shape $\tilde{X}_0$ is defined as a pre-selected landmark configuration representing the neutral pose or facial expression. Then, a transformation layer is built in this tangent space to increase global or local dissimilarities between class motions. This layer is followed by a CONV Block and a one-layer LSTM network, which learns the temporal dynamics of the landmark sequences. As output, a fully connected block yields the corresponding motion class. In the case of 3D landmark-based tasks, the CONV block consists of two 1D convolution layers followed by a pooling layer for dimensionality reduction. For end-to-end network training, we use the cross-entropy loss.

4.1 Optimization over rigid transformations

To optimize over rigid transformations, rotations are applied to individual 2D/3D landmark configurations across sequences within this layer, and are updated during training. Throughout, we use k to denote the dimension of the landmark points. Let $\tilde{Y}_i$ denote the i^{th} centered, unit norm landmark configuration in a sequence S, and $\hat{Y}_i$ its representative in the tangent space, reshaped from a $k(n-k)$ vector into a $k \times (n-1)$ matrix represented in the ambient coordinates. The transformation layer is performed on each sequence resulting in a hidden output h, given by:

$$h_i = O_i \hat{Y}_i \tag{4}$$

where $O_i \in SO(k)$. In the back-propagation phase, the gradient descent adapts the kernels O_i directly so that they may not lie in $SO(k)$. To ensure that the updated kernels lie in $SO(k)$, we propose a second variant of this layer, called angle-based, where the optimization is performed over the rotation angles (one rotation angle for $O \in SO(2)$ and three rotation angles for $O \in SO(3)$). Rotation matrices are then generated in the feed-forward pass.

Fig. 2 depicts the optimization over the rigid transformation layer on the human body 3D landmark-based representation and on the facial 2D landmark-based representation. We illustrate that this first category of optimization deals with a 2D/3D landmark configuration as a single rigid entity, i.e., the same rotation is applied to the entire landmark configuration.

4.2 Optimization over non-rigid transformations

The optimization over local transformations is performed by finding the best rotations of individual landmarks within each configuration. This more flexible modeling approach tends to improve performance on the motion analysis task. As before, let $\tilde{Y}_i$ denote the i^{th} centered, unit norm landmark configuration in a sequence S, $\hat{Y}_i$ its representative in the tangent space (reshaped from a $k(n-k)$ vector into a $k \times (n-1)$ matrix represented in the ambient coordinates), and $q_i^j \in \mathbb{R}^k$ the j^{th} landmark of $\hat{Y}_i$. The transformation layer is performed on each landmark resulting in a hidden output h, given by:

$$h_i = \{O_{i,j} q_i^j\}_{j=1}^n, \tag{5}$$

where $O_{i,j} \in SO(k)$. Similarly to the rigid transformation case, an angle-based optimization variant is proposed to ensure that each $O_{i,j}$ is a rotation matrix. In Section 5.1.4, we perform a study that compares the two variants for optimization over rigid and non-rigid transformations: 1) the variant that allows the network to use general kernels as $k \times k$ matrices (not necessarily elements of $SO(k)$), and 2) the angle-based approach that constrains the network to allow rotation matrices only.

Fig. 3 depicts the optimization over the non-rigid transformation layer on the human body 3D landmark-based representation and on the facial 2D landmark-based representation. We illustrate that, in contrast to optimization over rigid transformations, this

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3291663

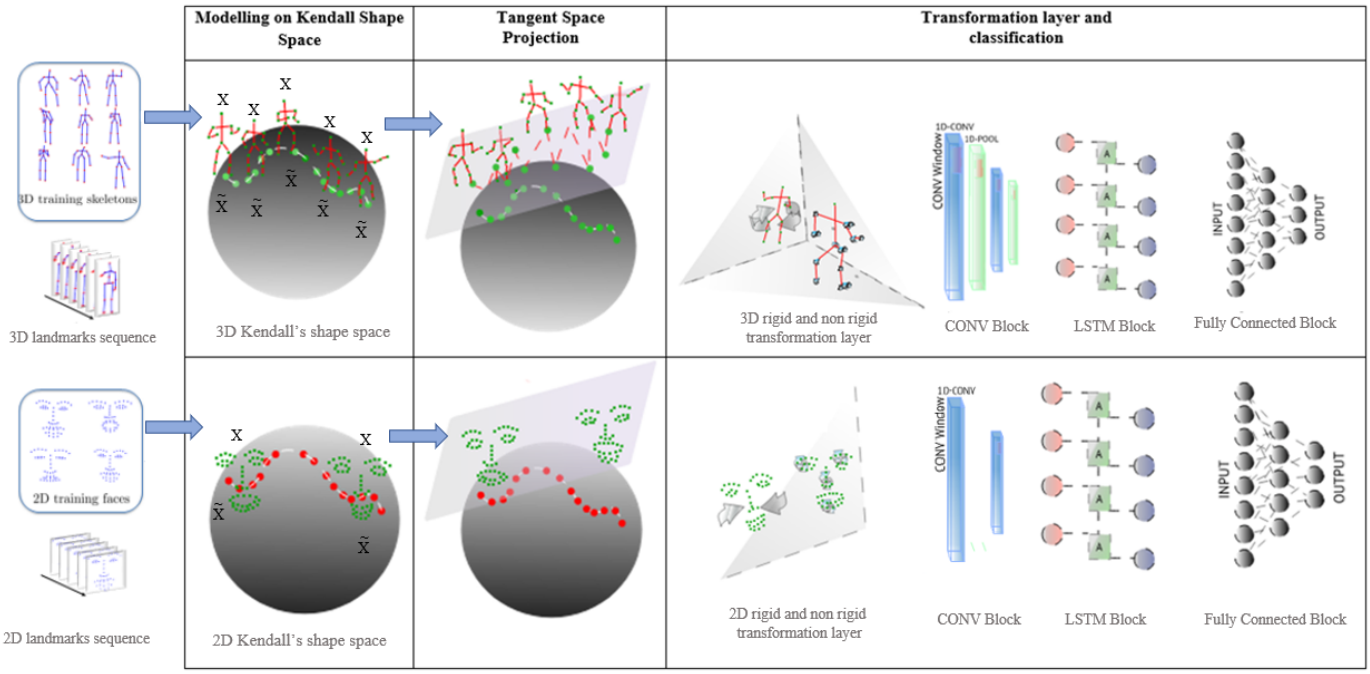

Fig. 1. Illustration of the full architecture of the proposed KShapeNet framework and its different blocks. 1) Representation of the input 2D/3D landmark sequences as trajectories on Kendall's shape space. 2) Projection of the sequences onto the tangent space at a landmark configuration corresponding to a neutral pose or face expression. 3) Deep learning architecture embedding the rigid and non-rigid transformation layers.

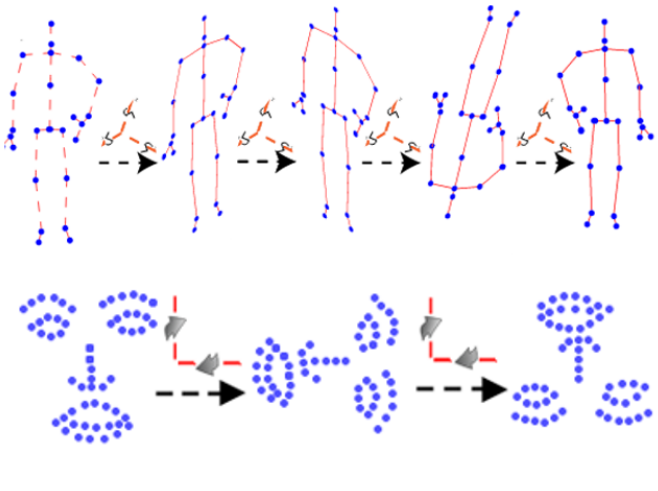

Fig. 2. Illustration of optimization over rigid transformations: rotations of the entire landmark configuration are applied during training.

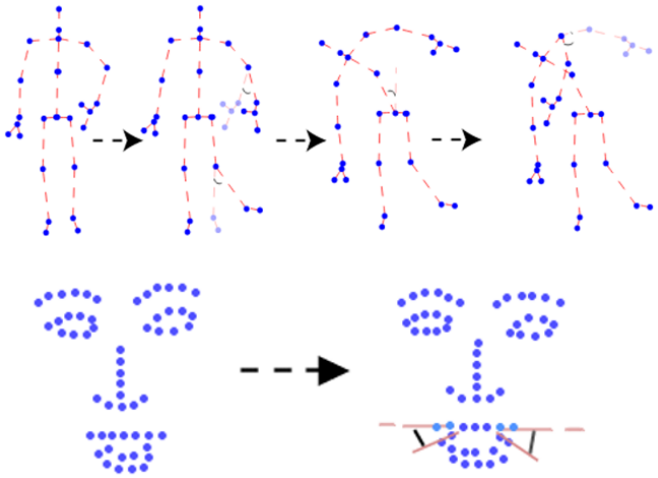

Fig. 3. Illustration of optimization over non-rigid transformations: rotations are applied to individual landmarks during training.

second category of optimization treats each 2D/3D landmark configuration as a non-rigid object, i.e., different rotations are applied to each individual landmark within each configuration.

5 Experimental Results

5.1 3D human action recognition

In Section 5.1.1, we first describe the datasets and experimental settings used to validate our architecture for the action recognition task. For the demonstration of KShapeNet efficiency, an ablation study is presented in Section 5.1.2 with a discussion of the impact of intermediate layers, i.e., the transformation layer and logarithmic map layer (projection to tangent space). Then, in Section 5.1.3, we compare the performance of action recognition based on the KShapeNet architecture to state-of-the-art approaches on the same datasets. We conclude in Section 5.1.4 with a comparison and discussion of pre-processing techniques, effects of nuisance variation, different variants of the transformation and projection on tangent space layers. The implementation code for KShapeNet will be publicly released upon acceptance of the paper.

5.1.1 Datasets and implementation settings

We evaluate the effectiveness of KShapeNet on two large scale state-of-the-art datasets, NTU-RGB+D and NTU-RGB+D120.

NTU-RGB+D (NTU) [69] is one of the largest 3D human action recognition datasets. It consists of 56,000 action clips of 60 classes. 40 participants were asked to perform these actions in a

constrained lab environment, with three camera views recorded simultaneously. Kinect sensors estimated and recorded the 3D coordinates of 25 joints in the 3D camera's coordinate system. For standard assessment, we utilize two state-of-the-art protocols: cross-subject (CS) and cross-view (CV). In the CS protocol, the 40 subjects are split into training and test sets (20 subjects each) made up of 40,320 and 16,560 samples (sequences), respectively. In the CV protocol, we select the samples from cameras 2 and 3 for training, and the samples from camera 1 for testing. Thus, the training set consists of the front and two side views of the actions, while the test set consists of the left and right 45 degree views. For this protocol, the training and test sets contain 37,920 and 18,960 samples (sequences), respectively.

NTU120-RGB+D (NTU120) [70] is an extension of NTU60. It is the largest RGB+D dataset for 3D action recognition with 114,480 skeleton sequences. It contains 120 action classes performed by 106 distinct human subjects. NTU120 was built using 32 collection setups; over different setups, location and background were changed. Specifically, in each setup, three cameras were used at the same time to capture three different horizontal views for the same action sample. For this dataset, the two protocols used for evaluation are cross-subject (CS) and cross-setup (Cset). For the CS setting, half of the 106 subjects are used for training and the rest for testing. For the Cset evaluation, all of the samples with even collection setup IDs are chosen for training, and those with odd setup IDs are used for testing, i.e., 16 setups are used for training and the other 16 setups are reserved for testing.

For the KShapeNet implementation, we set the number of frames to $p = 100$, and the batch size to 64 for the NTU dataset and 32 for the NTU120 dataset. To estimate the model's parameters, we use cross-entropy loss and set the number of epochs to 30. The Adam optimizer is adapted to train the network, and the initial learning rate is fixed to 1×10^{-4} for both datasets. For training, we used a machine with a processor speed of 3.40 GHz, memory of 32 GB and an NVIDIA GTX 1070 Ti GPU.

5.1.2 *Ablation study*

To validate the effectiveness of the proposed framework and highlight the impact of each processing block, we performed an ablation study by gradually adding 1) the logarithmic map block (projection to tangent space), and 2) the transformation layer.

Table 1 reports the results of this study on the NTU and NTU120 datasets. In the first row, labeled "Baseline", we report the results generated by the deep network (CNN-LSTM used in KShapeNet) using input data represented directly on Kendall's pre-shape space (without moving to the linear tangent space), and without the optimization over rigid or non-rigid transformations. The baseline architecture provides satisfactory results. However, they are not competitive with respect to those produced by state-of-the-art approaches.

Dataset	NTU-RGB+D		NTU-RGB+D120	
Protocol	CS	CV	CS	Cset
Baseline	85.1%	91.2%	56.0%	63.5%
Transformation layer only	89.6%	91.5%	57.2%	63.8%
Logarithmic map only	94.1%	95.5%	63.9%	65.3%
Proposed (KShapeNet)	**97.0%**	**98.5%**	**90.6%**	**86.7%**

TABLE 1
Ablation study results on the NTU and NTU120 datasets.

The second row of Table 1 depicts the results achieved by adding the transformation layer to the baseline architecture. The transformation layer adopted here considers optimization over non-rigid transformations using the angle-based variant. Further discussion about the choice of this configuration is presented in Section 5.1.4. Compared to the "Baseline" results, the transformation layer improves recognition performance by 4.5% for CS and 0.3% for CV on the NTU dataset, and by 1.2% for CS and 0.3% for Cset on the NTU120 dataset. As explained in Section 4, this configuration of the transformation layer optimizes over non-rigid transformations, hence urging the network to find the best local rotations that are applied to the skeleton joints within each landmark configuration; this justifies the improvement in action recognition accuracy.

In the third row of Table 1, we present the results obtained by only adding the projection to tangent space block, via the logarithmic map, to the baseline model. Linearization via tangent space projection provides significant improvements in recognition performance, increasing from 85.1% to 94.1% for the CS protocol on the NTU dataset. The increase in accuracy is due to a new skeleton landmark configuration representation in the Euclidean tangent space, allowing for the definition of a linear metric between skeleton landmark configuration shapes.

In the fourth row of Table 1, we report the final results produced by the KShapeNet framework, including both the logarithmic map block and the non-rigid transformation layer. KShapeNet results in a significant improvement over the baseline model, and most importantly, further increases recognition accuracy over the two models with individually added components (logarithmic map block or transformation layer). The combination of both components empowers the network to properly discriminate action classes. For instance, for the CS protocol on the NTU dataset, the accuracy increase due to the additional transformation layer was only 4.5% and the increase due to the logarithmic map block was only 9%. However, the addition of both components increased recognition accuracy by more than 11%. Accordingly, we conclude that the efficiency of KShapeNet is not only due to the advanced feature extraction capacity of the CNN-LSTM network, but equally due to the convenient data representation of skeleton landmark configuration shapes in the linear tangent space, and the optimization over local rotations.

5.1.3 *Comparison to state-of-the-art approaches*

In this section, we compare the performance of the proposed framework to state-of-the-art approaches on the two datasets, NTU and NTU120. Table 2 reports recognition results of state-of-the-art approaches on the NTU dataset, and compares them to the result generated by KShapeNet. In this table, we distinguish between three classes of action recognition methods: deep learning methods, Riemannian methods, and hybrid (deep Riemannian) methods; our framework, KShapeNet, falls into the third category. The results demonstrate that KShapeNet consistently outperforms deep learning (leveraging CNNs and RNNs), Riemannian and even hybrid approaches. Indeed, our method outperforms the best of these state-of-the-art approaches by 7.3% and 0.1% on the CS and CV settings, respectively. Comparing to the hybrid method presented in [13], which incorporates the Lie group structure into a deep network architecture using rotation mapping layers, our approach increases recognition accuracy by more than 35%.

Table 3 compares KShapeNet recognition accuracy to state-of-the-art approaches on the NTU120 dataset. KShapeNet achieves

competitive recognition results under the Cset protocol, and outperforms the top competitor (MS-G3D Net) under the CS protocol by 3.7%.

5.1.4 Additional studies

Next, we present intermediate experiments that were performed during the design of KShapeNet. In particular, we discuss the different configurations that were tested in terms of data preprocessing, and the variants of the transformation layer and the logarithmic map block.

Comparison of preprocessing techniques: We used the code of Maosen et al. [38] to generate input data for our algorithm, which is composed of three main steps: 1) extraction of skeleton landmark configurations across frames, 2) extraction of joint coordinates for each configuration, and 3) splitting of sequences into training and test sets for the different protocols. As an additional data processing step, we interpolated the sequences, using cubic splines, to estimate equally-spaced skeleton landmark configuration trajectories, with constant time change between frames. For comparison, we tested the network by zero padding the missing frames. Since this operation results in frames that contain "wrong" data, the network is misled during the learning stage and recognition performance deteriorates significantly. Table 4 reports recognition results, on the NTU dataset, obtained with zero padding and with cubic spline interpolation.

Effects of nuisance variation: As detailed previously, the representation in Kendall's shape space consists of filtering out the three sources of nuisance variation: translation, scale and rotation. In this section, we report recognition performance when only a subset of the nuisance variations is accounted for.

Table 5 contains recognition accuracies when two out of the three nuisance variations are filtered out from the representation space, for both the NTU and NTU120 datasets. The first row in Table 5 reports the accuracy obtained when translation variation is retained while filtering out scale and rotation. The second row reports the accuracy when keeping the initial rotations of the landmark configurations while filtering out scale and translation. Finally, the third row reports the accuracy when keeping scale while removing translation and rotation. From these results, it is clear that accounting for nuisance variation due to rotation and translation is an important aspect of the proposed approach.

Comparison of transformation layer variants: Table 6 presents a comparison of recognition results computed using the four different variants of the transformation layer, for the NTU and NTU120 datasets. Each row in the table refers to one of the four variants: 1) optimization over rigid rotations using the matrix-based variant (Rigid Matrix), 2) optimization over rigid rotations using the angle-based variant (Rigid Angle), 3) optimization over non-rigid rotations using the matrix-based variant (Non-rigid Matrix), and 4) optimization over non-rigid rotations using the angle-based variant (Non-rigid Angle).

At a global level, we notice that each version of the transformation layer preserves state-of-the-art results on the NTU and NTU120 datasets. Further, performance is generally better on the CV protocol than the CS protocol for NTU, and on the CS protocol than the Cset protocol for NTU120. At a granular level, we highlight two different behaviors of the optimization over rigid transformations and the optimization over non-rigid transformations, with regards to the two different variants: rotation matrix-based and angle-based. On the one hand, the rotation matrix-based variant, which gives the network the liberty to optimize matrix coefficients without any constraints (updated matrices may not be in $SO(3)$), yields better results for the optimization over rigid transformations than for the optimization over non-rigid ones. On the other hand, the angle-based variant, which only updates the angles resulting in elements of $SO(3)$, performs worse for rigid transformations than non-rigid ones.

Rigid transformations, i.e., rotations of the entire skeleton landmark configuration, are characterized by preserving the skeleton landmark configuration's shape, distance and angle properties, i.e., all joints move in the same direction by the same amount. We argue that, for this reason, the rotation matrix-based variant is more adequate for optimization over such transformations. In other words, the rigid transformation is not subject to shape and angle variations, and the network tends to perceive the transformations applied to the skeleton landmark configuration as a one entity operation. Therefore, it is more efficient to allow the network to freely optimize over matrices during the back forward phase without the orthogonality constraint. As a result of the non-rigid transformations, i.e., different rotations applied to all of the joints, the shape and angle properties of the skeleton landmark configurations are not preserved at each pass. Beyond the first feed forward pass, the network will alter the representation of each sequence. Thus, for the optimization over non-rigid transformations, it is more convenient to constrain the network to allow rotations only. The rotation matrices are generated based on updated rotation angles, always resulting in elements of $SO(3)$.

In light of these results, we chose to optimize over non-rigid transformations using the angle-based variant for the final configuration of KShapeNet. This allows for flexible modeling of inter-joint transformations; the corresponding recognition results are highlighted in bold in Table 6.

Study of the combination of rigid and non-rigid transformation layers: Up to this point, the rigid and non-rigid transformation layers were considered separately. However, they are not complementary. A rigid transformation, inducing a global rotation of a 3D landmark configuration, is a special case of a non-rigid transformation where all local joints are rotated in the same exact manner. The present section aims to validate this and experimentally investigate the non-complementary nature of the two transformation layers. To do this for each of the two variants of the transformation layers (matrix-based and angle-based), we merge the rigid and non-rigid transformations by implementing two subnetworks, each applying a specific transformation (rigid or non-rigid) to the input landmark configurations, followed by the convolution and LSTM layers. We then concatenate the output features before using them as input into the fully connected layer. The combination of features is applied on the linear functions of the two subnetworks.

Table 7 summarizes the recognition results of this study, for the NTU and NTU120 datasets, and compares them to the KShapenet accuracy. Comparing recognition accuracies reported in Table 7 to those in Table 6, we observe that, for the matrix-based variant, the combination approach generates a slight improvement over rigid transformations alone. For the angle-based variant, the combination approach decreases accuracy compared to rigid and non-rigid transformations applied separately. In all cases, the proposed KShapeNet architecture, based on the non-rigid angle-based transformation layer, yields the best recognition accuracy.

Comparison of different methods for projection to tangent space: As another intermediate experiment, we tested two approaches for the projection to tangent space block. The first

NTU-RGB+D Dataset		
Deep learning methods	Cross Subject	Cross View
Directed Graph Neural Networks [71]	89.9%	96.1%
Two stream adaptive GCN [72]	88.5%	95.1%
LSTM based RNN [34]	89.2%	95.0%
AGC-LSTM(Joints&Part) [73]	89.2%	95.0%
Riemannian methods	Cross Subject	Cross View
Lie Group [74]	50.1%	52.8%
Intrinsic SCDL [56]	73.89%	82.95%
Deep Riemannian methods	Cross Subject	Cross View
Deep learning on $SO(3)^n$ [13]	61.37%	66.95%
Proposed (KShapeNet)	**97.0%**	**98.5%**

TABLE 2
Comparison of KShapeNet to state-of-the-art approaches on the NTU dataset.

NTU120-RGB+D Dataset		
Method	Cross Subject	Cross Setup
Tree Structure + CNN[75]	67.9%	62.8%
SkeleMotion[76]	67.7%	66.9%
Body Pose Evolution Map[77]	64.6%	66.9%
MS-G3D Net[78]	86.9%	88.4%
Proposed (KShapeNet)	**90.6%**	**86.7%**

TABLE 3
Comparison of KShapeNet to state-of-the-art approaches on the NTU120 dataset.

NTU-RGB+D Dataset		
Protocol	CS	CV
Zero padding	81.3%	85.1%
Proposed (KShapeNet: Interpolation)	97.0%	98.5%

TABLE 4
Comparison of recognition accuracy, on the NTU dataset, when data was pre-processed by zero padding and cubic spline interpolation.

Dataset	NTU-RGB+D		NTU120-RGB+D	
Protocol	CS	CV	CS	Cset
Scale and rotation	89.6%	96.6%	82.1%	82.5%
Scale and translation	75.2%	74.3%	70.6%	63.1%
Rotation and translation	92.3%	95.2%	89.7%	84.3%
Proposed (KShapeNet: three variabilities removed)	97.0%	98.5%	90.6%	86.7%

TABLE 5
Impact on recognition accuracy when accounting for only a subset of nuisance variation when computing Kendall's shape space coordinates of the landmark configurations. Each row lists the two sources of nuisance variability that were removed from the representation space.

Dataset	NTU-RGB+D		NTU120-RGB+D	
Protocol	CS	CV	CS	Cset
Rigid matrix	97.0%	97.1%	90.2%	85.9%
Rigid angle	96.9%	96.3%	89.1%	84.9%
Non-rigid matrix	96.8%	96.9%	90.6%	84.3%
Proposed (KShapeNet: Non-rigid angle)	**97.0%**	**98.5%**	**90.6%**	**86.7%**

TABLE 6
Comparison of recognition accuracy based on the four different variants of the transformation layer.

Dataset	NTU-RGB+D		NTU-RGB+D120	
Protocol	CS	CV	CS	Cset
Matrix-based combination	96.9%	98.3%	90.4%	86.6%
Angle-based combination	96.5%	96.1%	88.9%	84.1%
Proposed (KShapeNet: Non-rigid angle)	97.0%	98.5%	90.6%	86.7%

TABLE 7
Impact of the combination of rigid and non-rigid transformation layers.

approach uses the logarithmic map to project all landmark configuration sequences to a single tangent space defined at a common reference configuration. In this variant, the distances between landmark configuration shapes computed in the tangent space are different than those computed directly on Kendall's pre-shape space (the only distances that are preserved after the projection are those from the reference to each projected shape). The issue is exacerbated when projecting landmark configuration shapes that are far away from the reference configuration.

To push the capabilities of our model, we next tried to incorporate parallel transport (PT) (refer to Section 3) as an alternative approach to map the landmark configuration sequences from the pre-shape space to the tangent space. In this approach, we first computed the shooting vectors between each consecutive frame within each sequence (using the logarithmic map). We then used PT to map these shooting vectors to the tangent space at the reference landmark configuration.

Table 8 presents the results of applying the one-shot logarithmic map and the PT approach, on the NTU dataset. Theoretically, PT should perform better than the direct projection to a tangent space at a reference landmark configuration since it remedies the distortion issues mentioned earlier. Nevertheless, as shown in Table 8, the simpler approach, logarithmic map, paradoxically tends to outperform the PT approach based on overall accuracy. In our implementation, the PT-based mapping to the tangent space iterations were not performed along the whole geodesic path, because this would have been computationally expensive. This in part justifies the better performance of the simple logarithmic map projection at a common reference point over the more complicated

Dataset	NTU-RGB+D	
Protocol	CS	CV
Parallel transport	96.8%	96.7%
Proposed (KShapeNet: Logarithmic map w.r.t. a reference frame)	97.0%	98.5%

TABLE 8
Comparison of recognition accuracy when projecting to tangent space using parallel transport and the logarithmic map at a reference landmark configuration.

PT approach.

Convergence of loss function: Lastly, we study the speed of convergence in two scenarios: 1) using raw input data, i.e., without translation and scale normalization, projection to the tangent space or transformation layer, and 2) using input data mapped onto the tangent space via Kendall's shape space coordinates with a non-rigid angle-based transformation layer, i.e., the KShapeNet implementation. Figure 4 shows the (a) training and (b) test loss, as a function of epochs, for scenario 1) (green) and scenario 2) (red) for the NTU dataset under the CS protocol. Under both scenarios, we observe convergence to a final error value. However, convergence under scenario 2) (KShapeNet) is faster and to a lower error value, resulting in higher recognition accuracy.

To summarize, at the end of the various experiments, we decided to adopt the following configurations for KShapeNet: projection on the tangent space using the logarithmic map with respect to a reference frame and optimization over non-rigid transformations using the angle-based variant (corresponding results are cited in Table 2 and Table 3). For precision, since the first frame in all of the landmark configuration sequences in the two datasets is neutral, i.e., they are very close to each other on Kendall's pre-shape space, we alternatively considered an approximation of the reference frame with the first frame of each sequence. In other words, we chose to map each sequence to the tangent space defined at the landmark configuration corresponding to its first frame, using the logarithmic map for this projection.

5.2 3D gait recognition

In this section, we use KShapeNet for the gait recognition task. As described in Section 4, input gait landmark configuration sequences are first represented as trajectories on C, after which each landmark configuration $\tilde{X}$ is mapped to a common tangent space $T_{\tilde{X}_0}(C)$ at a reference $\tilde{X}_0$, which is defined as a pre-selected configuration representing the neutral pose. Then, the transformation layer is built in this tangent space and followed by a CONV Block and a one-layer LSTM network. As output, a fully connected block yields the corresponding class. We train the network for 50 epochs and use cross-entropy as the training loss.

5.2.1 *Dataset and implementation settings*

For gait recognition, we evaluate the KShapenet framework on the CMU Mocap 3D gait dataset.

CMU Mocap is a database from the CMU Graphics Lab that contains multiple human motion sequences such as playing, running and walking. Motions of 144 subjects were recorded with an optical marker-based Vicon system. Subjects wore a black jumpsuit with 41 markers taped to it. The tracking space of $30m^2$ was surrounded by 12 cameras with a sampling rate of 120Hz at heights ranging from $2m$ to $4m$. In our work, we used the CMU Mocap 3D gait dataset extracted and released by Balazia et al. [79], [80], which includes 3843 gait cycles for 54 subjects.

Following the "Homogeneous" experimental setup described in [80], we used 10-fold cross validation to assess the recognition performance by dividing the evaluation set into one unlabeled fold as a test set and nine other labeled folds as a gallery set.

5.2.2 *Ablation Study*

To confirm the efficacy of the proposed KShapeNet architecture for 3D gait recognition, we conducted an ablation study on the CMU Mocap dataset in the same fashion as described in Section 5.1.2. Table 9 reports the recognition results. The adopted evaluation metric is average recognition accuracy across the different folds. These results, being in line with the results obtained for action recognition, consolidate our conclusion about the efficiency of KShapeNet due to the convenient data representation of 3D landmark configuration shapes in the linear tangent space, and the optimization over local rotation transformations.

Dataset	CMU Mocap
Baseline	82.30%
Transformation layer only	87.50%
Logarithmic map only	92.10%
Proposed (KShapeNet)	**96.02%**

TABLE 9
Ablation study results on the CMU Mocap dataset.

5.2.3 *Comparison to state-of-the-art approaches*

Next, we compare the performance of KShapenNet to state-of-the-art approaches on the CMU Mocap dataset. Table 10 reports these results. Again, KShapeNet consistently outperforms all of the competing deep learning-based architectures for gait recognition. Indeed, it outperforms the top competitor [81] by a small margin (96.02% vs. 95.97%).

5.2.4 *Additional studies*

As in the case of action recognition, we performed similar intermediate experiments using the CMU Mocap dataset for gait recognition. Here, we only focus on the comparison of the transformation layer variants and the comparison of the different methods of projection to the tangent space. Table 11 summarizes the results of these studies. The two columns in Table 11 correspond to the two approaches that can be used to map sequences to a tangent space, as discussed in Section 5.1.4: using the logarithmic map to a tangent space at a reference configuration and using PT. We can see that projection via the logarithmic map to a common reference frame outperforms projection via PT. Each of the four rows in Table 11 correspond to the four different transformation layer variants. We observe that the optimization over rigid transformations and the optimization over non-rigid transformations behave slightly differently compared to the action recognition application, with regards to the two different variants: rotation matrix-based and angle-based. The angle-based variant still performs better for non-rigid transformations than rigid ones. However, the matrix-based variant yields lower accuracy for optimization over rigid transformations. Similarly to action recognition, the non-rigid angle-based transformation layer outperforms the three other transformation variants.

As a final setting, we adopt the same configuration of KShapeNet: projection on the tangent space using the logarithmic

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3291663

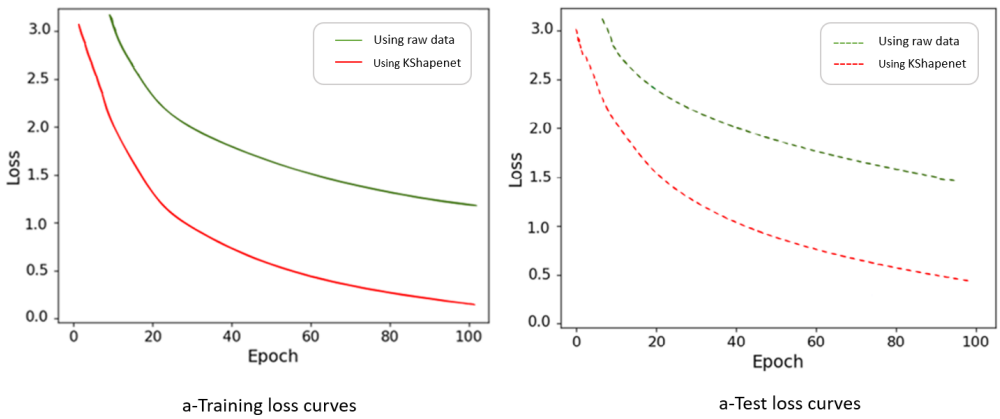

Fig. 4. Illustration of the (a) training and (b) test loss, as a function of epochs, for the NTU dataset under the CS protocol. Green: Baseline model for raw input data without projection to tangent space or transformation layers. Red: KShapeNet with non-rigid angle-based transformation layer.

Method	Year	Accuracy
Kwolek et al. [82]	2014	90.99%
Andersson et al. [83]	2015	77.87%
Balazia et al. [79] (PCA+LDA)	2016	83.14%
Balazia et al. [79] (MMC)	2016	91.02%
Hosni et al. [81]	2018	92.23%
Hosni et al. [81]	2020	95.97%
Proposed (KShapeNet)	2021	**96.02%**

TABLE 10
Comparison of recognition performance on the CMU Mocap gait dataset with respect to state-of-the-art.

	Logarithmic map w.r.t. a reference frame	Parallel transport
Rigid matrix	89.34%	72.82%
Rigid angle	91.36%	73.63%
Non-rigid matrix	90.23%	72.45%
Non-rigid angle	**96.02%**	77.56%

TABLE 11
Comparison of recognition accuracy based on different transformation layer variants and methods of projection to the tangent space on the CMU Mocap dataset.

map with respect to a reference frame and optimization over non-rigid transformations using the angle-based variant.

5.3 2D facial expression recognition

In this section, we use KShapeNet for the facial expression recognition task based on 2D landmark configurations. We use 50 epochs to train the network as in the previous section.

5.3.1 *Datasets and implementation settings*

For the facial expression recognition task, we evaluate the KShapeNet framework on two datasets: CK+ and Oulu-CASIA. As the first frame of all sequences in the two datasets is neutral, the first frame is considered as an approximation of the common reference frame.

Cohn-Kanade Extended (CK+) [84] is a dataset consisting of 327 video sequences of facial expressions performed by 118 subjects with seven emotion labels: anger, contempt, disgust, fear, happiness, sadness and surprise. Each sequence contains the two first temporal phases of the expression, i.e., neutral and onset (with apex frames). Each video shows a facial shift from the neutral expression to a targeted peak expression. The 118 subjects were divided into ten groups by ID in ascending order. Nine subsets were used for training the network, and the remaining subset was used for validation. This process is the same as the 10-fold cross validation protocol in [85].

Oulu-CASIA [86] is a dataset that includes 480 image sequences performed by 80 subjects. They are labeled with one of the six basic emotions (the same as in CK+, except contempt). Each sequence begins with a neutral expression and ends with the expression apex. The imaging hardware worked at the rate of 25 frames per second and the image resolution was 320×240 pixels.

5.3.2 *Ablation study*

To confirm the efficacy of the proposed KShapeNet architecture for 2D facial expression recognition, we conducted an ablation study on the CK+ and Oulu-CASIA datasets in the same fashion as described in Section 5.1.2. Table 12 reports the recognition results. The conclusions here are very similar to the previous two ablation studies that were conducted in the context of activity and gait recognition.

Dataset	CK+	Oulu-CASIA
Baseline	83.64%	74.15%
Transformation layer only	87.58%	76.04%
Logarithmic map only	93.84%	79.36%
Proposed (KShapeNet)	**96.91%**	**82.70%**

TABLE 12
Ablation study results on the CK+ and Oulu-CASIA datasets.

5.3.3 *Comparison to state-of-the-art approaches*

We compare the recognition performance of the KShapeNet architecture to state-of-the-art approaches for the facial expression recognition task on the two datasets, CK+ and Oulu-CASIA, in Table 13. While it does not yield the best performance on either of the two datasets, the KShapeNet framework provides very competitive results.

5.3.4 *Additional studies*

As for the action recognition and gait recognition applications, we report results of intermediate experiments for the facial expression

Method	CK+	Oulu-CASIA
(G) DTGN [57]	92.35%	74.17%
(A+G) DTAGN [57]	**97.25%**	81.46%
(R) Gram matrix trajectories [59]	**96.87%**	**83.13%**
(R) Extrinsic SCDL (SVM) [56]	95.62%	77.06%
(R) Extrinsic SCDL (Bi-LSTM) [56]	95.73%	73.09%
Proposed (KShapeNet)	**96.91%**	**82.70%**

TABLE 13
Comparison of KShapeNet to state-of-the-art approaches on the CK+ and Oulu-CASIA datasets. (A) Appearance-based approach. (G): Geometric approach. (R): Riemannian approach.

Dataset	CK+	Oulu-CASIA
Rigid matrix	88.98%	78.84%
Rigid angle	89.01%	79.15%
Non-rigid matrix	91.40%	80.78%
Non-Rigid angle	**96.91%**	82.70%

TABLE 14
Comparison of recognition accuracy based on different transformation layer variants for the CK+ and Oulu-CASIA datasets.

recognition task based on the CK+ and Oulu-CASIA datasets. Here, we focus only on comparing the four different transformation layer variants and use the logarithmic map for projection to a tangent space at the reference configuration. Results for the CK+ and Oulu-CASIA datasets are summarized in Table 14.

We observe that the optimization over rigid transformations and the optimization over non-rigid transformations behave the same way as for gait recognition, but slightly differently compared to action recognition, with regards to the two different variants: rotation matrix-based and angle-based. The angle-based variant still performs better for non-rigid transformations than rigid ones. However, the matrix-based variant yields lower results for the optimization over rigid transformations. Similarly to action recognition and gait recognition, the non-rigid angle-based transformation layer outperforms the three other transformation variants.

As a final setting, we adopt the same configuration for KShapeNet: projection on the tangent space using the logarithmic map with respect to a reference frame and optimization over non-rigid transformations using the angle-based variant.

6 Conclusion

In this paper, we proposed a geometric deep architecture, KShapeNet, for human motion analysis based on modeling landmark sequences on Kendall's shape space. As part of the framework, we introduced a novel transformation layer to increase global or local dissimilarities between different types of motion. In the transformation layer, we optimize over rigid or non-rigid transformations. In addition, we explored the use of two optimization variants: 1) rotation matrix-based, and 2) angle-based. We showed that the matrix-based variant yields better performance when optimizing over rigid transformations, while the second yields better performance when optimizing over non-rigid transformations. Extensive experiments on challenging datasets, two for action recognition, one for gait recognition and two for facial expression recognition, demonstrate that the proposed framework performs very well compared to state-of-the-art approaches.

As future work, we will use the landmark shape sequence representation and optimization methods proposed in this paper to develop an unsupervised system for motion analysis, which will enable us to move beyond tasks which require labeled data.

Acknowledgment

This research was supported in part by NIH R37-CA214955, NSF CCF-1740761, NSF CCF-1839252 and NSF DMS-2015226 (to SK).

References

[1] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multim.*, vol. 19, no. 2, pp. 4–10, 2012.
[2] J. Lin, C. Gan, and S. Han, "Temporal shift module for efficient video understanding," *CoRR*, vol. abs/1811.08383, 2018.
[3] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.
[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, 2014, pp. 568–576.
[5] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.
[6] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," pp. 5669–5678, 2017.
[7] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018.
[8] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, 2014, pp. 1859–1866.
[9] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *13th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE Computer Society, 2018, pp. 59–66.
[10] Z. Sun, J. Liu, Q. Ke, H. Rahmani, M. Bennamoun, and G. Wang, "Human action recognition from various data modalities: A review," *CoRR*, vol. abs/2012.11866, 2020.
[11] D. G. Kendall, "Shape manifolds, procrustean metrics, and complex projective spaces," *Bulletin of the London mathematical society*, vol. 16, no. 2, pp. 81–121, 1984.
[12] Z. Huang, J. Wu, and L. V. Gool, "Building deep networks on grassmann manifolds," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 2018, pp. 3279–3286.
[13] Z. Huang, C. Wan, T. Probst, and L. V. Gool, "Deep learning on lie groups for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017, pp. 1243–1252.
[14] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.
[15] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst, "Shapenet: Convolutional neural networks on non-euclidean manifolds," Tech. Rep., 2015.
[16] R. Friji, H. Drira, and F. Chaieb, "Geometric deep learning on skeleton sequences for 2d/3d action recognition," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, Volume 5: VISAPP*, 2020, pp. 196–204.
[17] N. Hosni and B. B. Amor, "A geometric convnet on 3d shape manifold for gait recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, 2020, pp. 3725–3734.
[18] R. Friji, H. Drira, F. Chaieb, H. Kchok, and S. Kurtek, "Geometric deep neural network using rigid and non-rigid transformations for human action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 611–12 620.
[19] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *IJCAI, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. IJCAI/AAAI, 2013, pp. 2466–2472.
[20] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2014, pp. 588–595.
[21] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "RNN fisher vectors for action recognition and image annotation," in *European Conference on Computer Vision - ECCV*, vol. 9910, 2016, pp. 833–850.

[22] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: pose-based CNN features for action recognition," in *IEEE International Conference on Computer Vision, ICCV*, 2015, pp. 3218–3226.

[23] Y. Kong, L. Li, K. Zhang, Q. Ni, and J. Han, "Attention module-based spatial-temporal graph convolutional networks for skeleton-based action recognition," *J. Electronic Imaging*, vol. 28, no. 04, p. 043032, 2019.

[24] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 3697–3704.

[25] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," in *IEEE Winter Conference on Applications of Computer Vision, WACV*, 2017, pp. 148–157.

[26] W. Li, L. Wen, M. Chang, S. Lim, and S. Lyu, "Adaptive RNN tree for large-scale human action recognition," in *IEEE International Conference on Computer Vision, ICCV*, 2017, pp. 1453–1461.

[27] R. Zhao, H. Ali, and P. V. der Smagt, "Two-stream RNN/CNN for action recognition in 3d videos," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 4260–4267.

[28] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3633–3642.

[29] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3d human action recognition," in *14th European Conference on Computer Vision - ECCV*, vol. 9907, 2016, pp. 816–833.

[30] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper RNN," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5457–5466.

[31] S. Yang, J. Yu, C. Hu, and H. Jiang, "Quasi-projective synchronization of fractional-order complex-valued recurrent neural networks," *Neural Networks*, vol. 104, pp. 104–113, 2018.

[32] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1044–1048, 2018.

[33] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowl. Based Syst.*, vol. 158, pp. 43–53, 2018.

[34] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, 2019.

[35] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *MMAsia'19: ACM Multimedia Asia*, 2019, pp. 31:1–31:6.

[36] A. H. Ruiz, L. Porzi, S. R. Bulò, and F. Moreno-Noguer, "3d cnns on distance matrices for human action recognition," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 1087–1095.

[37] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 2018, pp. 7444–7452.

[38] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019, pp. 3595–3603.

[39] I. Rida, N. Al-Máadeed, and S. Al-Máadeed, "Robust gait recognition: a comprehensive survey," *IET Biom.*, vol. 8, no. 1, pp. 14–28, 2019.

[40] A. M. Nambiar, A. Bernardino, and J. C. Nascimento, "Gait-based person re-identification: A survey," *ACM Comput. Surv.*, vol. 52, no. 2, pp. 33:1–33:34, 2019.

[41] J. M. Echterhoff, J. Haladjian, and B. Brügge, "Gait and jump classification in modern equestrian sports," in *ACM International Symposium on Wearable Computers, UbiComp*, 2018, pp. 88–91.

[42] A. Muro-de-la-Herran, B. Garcia-Zapirain, and A. Méndez-Zorrilla, "Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, 2014.

[43] T. T. Verlekar, P. aulo Lobato Correia, and L. ís Ducla Soares, "Using transfer learning for classification of gait pathologies," in *IEEE International Conference on Bioinformatics and Biomedicine, BIBM*. IEEE Computer Society, 2018, pp. 2376–2381.

[44] Y. Wang, J. Sun, J. Li, and D. Zhao, "Gait recognition based on 3d skeleton joints captured by kinect," in *IEEE International Conference on Image Processing, ICIP 2016, Phoenix*. IEEE, 2016, pp. 3151–3155.

[45] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *CoRR*, vol. abs/2102.09546, 2021.

[46] A. Ball, D. C. Rye, F. Ramos, and M. Velonaki, "Unsupervised clustering of people from 'skeleton' data," in *International Conference on Human-Robot Interaction, HRI'12*. ACM, 2012, pp. 225–226.

[47] M. Ding, W. Li, H. ongyan Wang, and Z. Zhao, "Human gait recognition based on multi-feature fusion and kinect sensor," in *12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI*. IEEE, 2019, pp. 1–6.

[48] V. O. Andersson and R. M. de Araújo, "Person identification using anthropometric and gait data from kinect sensor," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 425–431.

[49] B. Dikovski, G. Madjarov, and D. Gjorgjevikj, "Evaluation of different feature sets for gait recognition using skeletal data from kinect," in *37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO*, 2014, pp. 1304–1308.

[50] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien, "Gait recognition with kinect," in *1st International Workshop on Kinect in Pervasive Computing*, 2012, pp. 1—4.

[51] A. Sinha, K. Chakravarty, and B. Bhowmick, "Person identification using skeleton information from kinect," in *Intl. Conf. on Advances in CHI*, 2013, pp. 101—-108.

[52] N. Hosni, H. Drira, F. Chaieb, and B. B. Amor, "3d gait recognition based on functional PCA on kendall's shape space," in *24th International Conference on Pattern Recognition, ICPR*. IEEE Computer Society, 2018, pp. 2130–2135.

[53] S. Li and W. Deng, "Deep facial expression recognition: A survey," *CoRR*, vol. abs/1804.08348, 2018.

[54] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *12th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE Computer Society, 2017, pp. 118–126.

[55] A. S. Rokkones, M. Z. Uddin, and J. Tørresen, "Facial expression recognition using robust local directional strength pattern features and recurrent neural network," in *9th IEEE International Conference on Consumer Electronics, ICCE-Berlin*. IEEE, 2019, pp. 283–288.

[56] A. B. Tanfous, H. Drira, and B. B. Amor, "Sparse coding of shape trajectories for facial expression and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2594–2607, 2020.

[57] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, 2015, pp. 2983–2991.

[58] S. D. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *IEEE International Conference on Computer Vision Workshops, ICCV*. IEEE Computer Society, 2011, pp. 1642–1649.

[59] A. Kacem, M. Daoudi, B. B. Amor, and J. C. Á. Paiva, "A novel space-time representation on the positive semidefinite cone for facial expression recognition," in *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, 2017, pp. 3199–3208.

[60] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 2016 ACM Conference on Multimedia Conference*. ACM, 2016, pp. 382–386.

[61] D. Y. Choi, D. H. Kim, and B. C. Song, "Recognizing fine facial micro-expressions using two-dimensional landmark feature," in *IEEE International Conference on Image Processing, ICIP*. IEEE, 2018, pp. 1962–1966.

[62] A. B. Tanfous, H. Drira, and B. B. Amor, "Sparse coding of shape trajectories for facial expression and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2594–2607, 2020.

[63] R. Anirudh, P. K. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of Riemannian trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 922–936, 2017.

[64] P. K. Turaga and R. Chellappa, "Locally time-invariant models of human activities using trajectories on the grassmannian," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2435–2441.

[65] Z. Huang and L. V. Gool, "A Riemannian network for SPD matrix learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

[66] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*. Wiley, 1998.

[67] X. Pennec, "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements," *J. Math. Imaging Vis.*, vol. 25, no. 1, pp. 127–154, 2006.

[68] S. Lang, *Fundamentals of Differential Geometry*. Springer, 1999.

[69] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016, pp. 1010–1019.

[70] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, 2020.

[71] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 2019, pp. 7912–7921.

[72] ——, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019, pp. 12 026–12 035.

[73] S. Chenyang, C. Wentao, W. Wei, W. Liang, and T. Tieniu, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

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

[75] C. Caetano, F. Brémond, and W. R. Schwartz, "Skeleton image representation for 3d action recognition based on tree structure and reference joints," in *32nd SIBGRAPI Conference on Graphics, Patterns and Images*, 2019, pp. 16–23.

[76] C. Caetano, J. S. de Souza, F. Brémond, J. A. dos Santos, and W. R. Schwartz, "Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition," pp. 1–8, 2019.

[77] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1159–1168.

[78] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Conference on Computer Vision and Pattern Recognition, CVPR*, 2020.

[79] M. Balazia and P. Sojka, "Learning robust features for gait recognition by maximum margin criterion," in *23rd International Conference on Pattern Recognition, ICPR*. IEEE, 2016, pp. 901–906.

[80] ——, "Gait recognition from motion capture data," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 14, no. 1s, pp. 22:1–22:18, 2018.

[81] N. Hosni and B. B. Amor, "A geometric convnet on 3d shape manifold for gait recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020, pp. 3725–3734.

[82] B. Kwolek, T. Krzeszowski, A. Michalczuk, and H. Josinski, "3d gait recognition using spatio-temporal motion descriptors," in *Intelligent Information and Database Systems - 6th Asian Conference, ACIIDS II*, 2014, pp. 595–604.

[83] V. O. Andersson and R. M. de Araújo, "Person identification using anthropometric and gait data from kinect sensor," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 425–431.

[84] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, 2010, pp. 94–101.

[85] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2014, pp. 1749–1756.

[86] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

Rasha Friji received her PhD in Computer Vision from National University of Computer Science ENSI, Manouba University Campus and she is member of CRISTAL Lab. She received her engineering degree and her M.S. degree from the National Engineering School of Tunis in Tunisia. She is an AI Research Engineer at Talan Innovation Factory,R&D Department of Talan Tunisia.

Faten Chaieb Faten Chaieb is Professor of Computer Science with Efrei Paris Engineering School and a member of Efrei Research Lab Paris Pantheon-Assas University since January 2020. She earned her Ph.D. in Computer Science from the National School of Computer Science (ENSI), University of Manouba (Tunisia) in 2001. She obtained her 'habilitation degree' from the same University in June 2016. Her research interests include pattern recognition, medical imaging, shape analysis and 3D coding. More recently, she has also become interested in geometric deep learning applied to 3D human behavior understanding.

Hassen Drira is Professor of Computer Science and image processing with university of Strasbourg and member of ICube UMR 7357, CNRS, France since September 2022. He earned his Ph.D. in Computer Science in University of Lille in 2011 and "Habilitation à diriger des recherches" from the latter university in 2020. His research interests include pattern recognition, shape analysis and computer vision. He has published several refereed journals and conference articles in these areas.

Sebastian Kurtek is Professor of Statistics at The Ohio State University. He received his Ph.D. in Biostatistics from Florida State University in 2012. His research interests include statistical shape analysis, functional data analysis and statistics on manifolds, among others. His research has been supported by the National Institutes of Health and the National Science Foundation.