electronics

MDPI

Article

Rate-Invariant Modeling in Lie Algebra for Activity Recognition

Malek Boujebli [1,*], Hassen Drira [2,3], Makram Mestiri [1] and Imed Riadh Farah [1]

[1] Software Engineering, Distributed Applications, Decision Systems and Intelligent Imaging Research Laboratory (RIADI), National School of Computer Science (ENSI), University of Manouba, Manouba 2010, Tunisia; mmestiri@gmail.com (M.M.); imed.riadh.farah@gmail.com (I.R.F.)
[2] IMT Lille Douai, Institut Mines-Télécom, Center for Digital Systems, F-59000 Lille, France; hassen.drira@imt-lille-douai.fr
[3] Univ. Lille, CNRS, Centrale Lille, Institut Mines-Télécom, UMR 9189-CRIStAL, F-59000 Lille, France
[*] Correspondence: mboujebli@gmail.com

Received: 21 September 2020; Accepted: 4 November 2020; Published: 10 November 2020

Abstract: Human activity recognition is one of the most challenging and active areas of research in the computer vision domain. However, designing automatic systems that are robust to significant variability due to object combinations and the high complexity of human motions are more challenging. In this paper, we propose to model the inter-frame rigid evolution of skeleton parts as the trajectory in the Lie group $SE(3) \times \ldots \times SE(3)$. The motion of the object is similarly modeled as an additional trajectory in the same manifold. The classification is performed based on a rate-invariant comparison of the resulting trajectories mapped to a vector space, the Lie algebra. Experimental results on three action and activity datasets show that the proposed method outperforms various state-of-the-art human activity recognition approaches.

Keywords: activity recognition; rate invariance; Lie group

1. Introduction

Human activity recognition has attracted many research groups in recent years due to its wide range of promising applications in different domains, like surveillance, video games, physical rehabilitation, etc. In order to develop systems for understanding human behavior, visual data form one of the most important cues compared to verbal or vocal communication data. Moreover, the introduction of low cost depth cameras with real-time capabilities, like the Microsoft Kinect, which provide in addition to the classical red-green-blue (RGB) image, a depth image, makes it possible to estimate in real time a 3D humanoid skeleton thanks to the work of Shotton et al. [1]. This type of data brings several advantages as it makes the background easy to remove and allows extracting and tracking the human body, thus capturing the human motion in each frame. Additionally, the 3D depth data are independent of the human appearance (texture), providing a more complete human silhouette relative to the silhouette information used in the past. Thus, new datasets with RGB-depth (RGBD) data have been collected, and many efforts have been made on human action recognition. However, human activity understanding is a more challenging problem due to the diversity and complexity of human behaviors, and less effort has been made by previous approaches. The interaction with objects creates an additional challenge for human activity recognition. Actually, during a human–object interaction scene, the hands may hold objects that are hardly detected or recognized due to heavy occlusions and appearance variations. The high level information of the objects is needed to recognize the human–object interaction. Taking a glance at the past skeleton-based human activity recognition approaches, we can distinguish two categories: the first family of approaches considers the skeleton data as body parts, and the second family considers them as a set of joints, as categorized by [2]. The scope of the paper is related to the first family of approaches that first consider the human skeleton as a connected set of rigid segments and either model the temporal evolution of individual body parts [3] or focus on connected pairs of body parts and model the temporal evolution of joint angles [4,5]. More recently, Vemulapalli et al. [2] proposed to model a skeleton by all the possible rigid transformations between its segments. In other words, for each skeleton, hundreds of rotations and translations (let us assume L) are computed between all the skeleton segments to yield L points on the special euclidean group $SE(3)$. The evolution of the skeleton along frames generates a trajectory in $SE(3)^L$. The trajectories are later on mapped to the Lie algebra (the tangent space on the identity point of the special euclidean group). The main limitation of this approach is the distortions caused by this mapping especially for points far from the identity element. The authors proposed an improvement of this method by the rolling-based approach in order to minimize the distortions in the tangent space (Lie algebra) in [6]. In this paper, we propose to investigate transformations of each skeleton part along frames in the Lie group and not within the same frame as [2,6]. Compared to [2] and [6], the proposed model represents three main advantages:

- For a skeleton with n joints, we manage a trajectory in $SE(3)^{n-1}$ compared to a trajectory in a much bigger space $SE(3)^{n\times(n-1)}$ as modeled in [2,6]. This makes the proposed approach faster than that of [2,6].
- The mapping into the tangent space on the identity element does not cause distortions in the proposed approach as the transformations are considered for the same body segment across frames, and thus, the resulting points on the special euclidean group are close to the identity element.
- We model the object within the human–object interaction and present results on datasets including human–object interaction, which was not the handled in [2,6].

The main contributions of this work are:

- We perform a spatio-temporal modeling of skeleton sequences as trajectories on the special euclidean group.
- The rigid transformations of the object are modeled as an additional trajectory in the same manifold, while in [7], only the joint-based approaches were proposed.
- An elastic metric of the trajectories is proposed to model the time independently of the execution rate.
- Exhaustive experiments and comparative studies are presented on three benchmarks: a benchmark for action without objects (MSR-Action dataset), a benchmark for actions with object interaction (SYSU3D Human-Object Interaction dataset), and a benchmark with a mixture of action and human–object interaction (MSR Daily Activity dataset).

The paper is organized as follow: We provide a brief review of the existing literature in Section 2 and discuss the spatio-temporal modeling in Section 3. Section 4 presents the rate invariance modeling and classification. We present our experimental results in Section 5 and conclude the paper in Section 6.

2. State-of-the-Art

Currently, the recognition of human activities has become more popular in the computer vision committee, and this interest is translated by many applications into more activities such as surveillance, video games, physical rehabilitation, etc. In this case, we can distinguish three emerging branches in the research on the recognition of activities: (1) depth-based representation, (2) skeleton-based representation, and (3) RGB-D-based development. In this section, we will go back to the existing work of recognizing human activities captured by depth cameras and describe in more detail the literature on the specific shared structures of learning for the recognition of activities.

2.1. Depth-Based Representation

In [8,9], descriptors previously designed for the deep RGB channel were generalized to describe the geometry of the shape and construct a depth-based representation. The limitation of the approach proposed in [8] is the sensitivity to the point of view as the sampling scheme depends on the view. Along similar lines, Oreifej and Liu [3] used the histogram of oriented gradients (HOG) to capture the distribution of the normal orientation of the surface in 4D space. Yang et al. [10] proposed to concatenate the normal vectors into a spatio-temporal sub-volume of depth together to capture more informative geometric clues. In the work of Lu et al. [11], to represent complex human activities involving human–object interactions without taking into account holistic human postures, discriminating local patterns were used, and also, the authors proposed to study the relationship between the sampled pixels in the actor and background regions. A common limitation of depth-based approaches is the view sensitivity and time consumption due to the heavy signature compared to skeleton-based approaches.

2.2. Skeleton-Based Representation

Human movements can be effectively captured by the positional dynamics of each skeletal joint [12–15], or the relationship between joint pairs [4,16], or even their combination [17–19]. In [8], a tool for monitoring the human skeleton (3D posture) in real time from an image at a single depth was developed. The existing skeleton-based human action recognition can be broadly grouped into two main categories: joint-based approaches and body part-based approaches. Joint-based approaches consider the human skeleton as a set of points, whereas body part-based approaches consider the human skeleton as a connected set of rigid segments between connected pairs of body parts. In [14], human skeletons were represented using the 3D joint locations, and a temporal hierarchy of co-variance descriptors was proposed to model joint trajectories. F.Lv and R.Nevatia in [15] proposed to use the hidden Markov models (HMMs) to represent the position of the joints. Devanne et al. [20] represented the 3D position evolution as a trajectory of movement. The problem of action recognition was then formulated as the problem of calculating the similarity between the shape of trajectories in a Riemannian manifold. Along similar lines, in these works [21,22] presented a Riemannian analysis of distance trajectories for real-time action recognition. In [23], the relative positions of pairwise articulations were used to represent the human skeleton, and the temporal evolution of this representation was modeled using a hierarchy of Fourier coefficients. X.Yang et al. in [16] proposed an effective method using the relative articular positions, temporal displacement of joints, and offset of the joints with respect to the initial frame.

The second category of skeleton-based approaches investigates the body parts. In [2], the human skeleton was represented by points in the Lie group $SE(3) \times \ldots \times SE(3)$, by explicitly modeling the 3D geometric relationships between various body parts within a frame using rotations and translations, then the human action was modeled as curves in this Lie group. The temporal evolution was handled by dynamic time warping (DTW). On the other hand, in [24], the human skeleton was hierarchically divided into smaller parts, and each part was represented using some bio-inspired features. Linear dynamic systems were used to model the temporal evolution of this part. Generally speaking, the joint-based method owns a faster calculation speed, while body the part-based method owns higher accuracy [25].

2.3. RGB-D-Based Development

The depth image is robust to lighting changes. However, it loses some useful information, such as texture context, which is essential to distinguish certain activities involving human-object interactions. Recently, several works showed that to improve the recognition of activities with object interactions, it is also necessary to merge RGB sequences with depth images [26–33]. For example, in the work of Zhao et al. [33], combined descriptors based on points of interest, extracted from RGB sequences and depth sequences, were brought together to perform the recognition. Liu and Shao [26] used a deep architecture to simultaneously merge RGB information and depth images; in [19], a set of random forests was used to merge spatio-temporal and human key articulations; Shahroudy et al. [27], used a structured density method by merging RGB information and skeleton indices, and in [29], the authors simply concatenated skeletal features and silhouette-based features to perform classification.

The review and analysis of current RGB-D action datasets revealed some limitations including size, applicability, availability of ground truth labels, and evaluation protocols. There is also the problem of dataset saturation, a phenomenon whereby the algorithms reported achieved a near-perfect performance.

3. Spatio-Temporal Modeling

3.1. Proposed Approach

In this work, we propose a framework for human activity recognition using the body part-based skeleton for action recognition and object detection and object tracking for human–object interaction recognition. Figure 1 summarizes the proposed approach: First, skeleton and object sequences are represented as trajectories in the Lie group, and these trajectories are then mapped into the Lie algebra, then to a Riemannian manifold to be compared in a rate-invariant way. In addition to distances to training trajectories, the output of the last layer of the neural network used for object detection is also used, in some scenarios, to build the final feature vector. The classification is therefore performed using the Hoeffding tree (“very fast decision trees (VFDT)”).

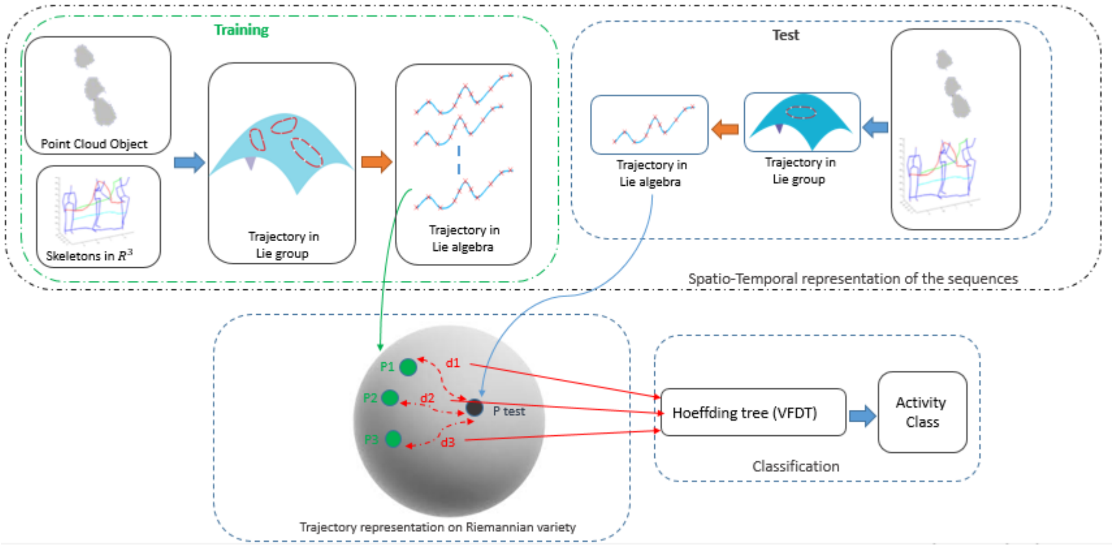


Figure 1. Overview of the proposed approach. VFDT, very fast decision trees.

Inspired by the work proposed in [2], which focused on the rigid transformations between different parts of the body within the same frame, we propose to model the evolution of the same part of the body (a segment) across frames by using the rotation and the translation necessary to transform the segment at frame t to the correspondent segment at frame $t+1$. This geometric representation of the rotation and translation of the rigid body in 3D space is part of the special euclidean group $SE(3)$ [34]. The evolution between two successive frames can be therefore modeled as a point in $SE(3)\times\ldots\times SE(3)$ $n-1$ times, where $n-1$ represents the number of body segments for a skeleton with n joints. A sequence of N frames is therefore represented by $N-1$ points in $SE(3)\times\ldots\times SE(3)$ ($n-1$ times) and can be modeled as a trajectory the in $SE(3)\times\ldots\times SE(3)$ ($n-1$ times) manifold. When an object is considered, an existing neural network is used for object detection in the first frame (RGB of the object in sequence i; frame j denoted by $OBJ-RGB(i)(j)$), then the object is tracked during the sequence (depth of the object in sequence i; frame j denoted by $OBJ-Depth(i)(j)$) using the iterative closest point (ICP) algorithm. The rigid deformations of the object across frames creates an additional trajectory in $SE(3)$ that is considered with the previous trajectory generated by the skeleton motion, to yield a final trajectory in $SE(3) \times \ldots \times SE(3)$ (n times). The next step is to map this trajectory to the corresponding Lie algebra $se(3) \times \ldots \times se(3)$, which is the tangent space at the identity element. The resulting trajectories lie in a euclidean space (Lie algebra) and incorporate the geometric deformations between body segments across frames. In order to compare their shapes independently of the execution rate, they are mapped to the shape space of continuous curves via the square root velocity manifold representation [35]. The classification is performed later using the Hoeffding tree (VFDT) [36] based on the elastic metric in the shape space.

3.2. Skeleton Motion Modeling

Firstly, we present the spatio-temporal modeling of the sequences. For this, we describe the geometric relation between the part of the body (denoted by *part*) at frame f_t and the same part in succession frame f_{t+1}. To do this, we use the rotation and translation required to move the current part to the position and orientation of the same part in the next frame, and we use the *procruste* function. This geometric transformation such as rotation and translation between two rigid body parts is a member of the special Euclidean group *SE(3)* [34] and defined by the following four by four matrix of the form:

$$P(R, \vec{d}) = \begin{bmatrix} R & \vec{d} \\ 0 & 1 \end{bmatrix} \in SE(3) \tag{1}$$

where $\vec{d} \in \mathbb{R}^3$ and $R \in \mathbb{R}^{3\times3}$ is a rotation matrix, which is a point on the special orthogonal group $SO(3)$.

This geometrical transformation between two parts of the rigid body with two successive frames is represented by a point in $SE(3)$. Obviously, all parts of the body are presented by a point of the Lie group $SE(3) \times \ldots \times SE(3)$, where × denotes the direct product between Lie groups. Therefore, the temporal transformation of the body parts can be modeled by a trajectory in the $SE(3) \times \ldots \times SE(3)$ Lie group, as depicted in Figure 2.

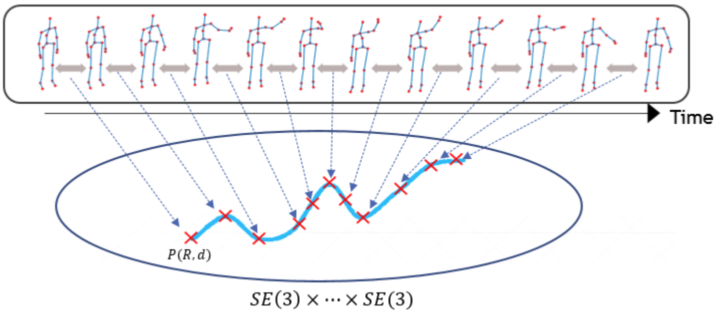


Figure 2. Action as a curve in the Lie group.

The Lie group identity element I_4 is defined by a four by four matrix. Mathematically, the tangent space to $SE(3)$ at the identity element is symbolized by $se(3)$, and it is considered to be the Lie algebra of $SE(3)$. This tangent space is a six-dimensional space constructed by matrices of the form:

$$B = \begin{bmatrix} U & \vec{w} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -u_3 & u_2 & w_1 \\ u_3 & 0 & -u_1 & w_2 \\ -u_2 & u_1 & 0 & w_3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \in se(3) \tag{2}$$

where $\vec{w} \in \mathbb{R}^3$ and $U \in \mathbb{R}^{3\times3}$ the skew-symmetric matrix. Thus, it can be presented as a six-dimensional vector:

$$vec(B) = \begin{bmatrix} u_1, u_2, u_3, w_1, w_2, w_3 \end{bmatrix} \tag{3}$$

The exponential map for $SE(3)$ is defined as $\exp_{SE(3)} : se(3) \rightarrow SE(3)$ and the inverse exponential map, defined as $\log_{SE(3)} : SE(3) \rightarrow se(3)$. Both are used to navigate between the manifold and the tangent space, respectively, given by:

$$\exp_{SE(3)}(B) = e^{B}, \log_{SE(3)}(P) = log(P) \tag{4}$$

where e and log denote the matrix exponential and logarithm, respectively.

The geometric transformation between all the parts of two successive frames f_t and f_{t+1} can be represented as:

$\delta(t) = (P_{f_t(1),f_{t+1}(1)}(t), P_{f_t(2),f_{t+1}(2)}(t) \ldots, P_{f_t(M),f_{t+1}(M)}(t)) \in SE(3) \times \ldots \times SE(3)$, where M is the number of body parts. Using this representation, a skeletal sequence describes an action as a curve in $SE(3) \times \ldots \times SE(3)$. One can not directly classify the action curves in the curved space $SE(3) \times \ldots \times SE(3)$, according to [2]. In addition, temporal modeling approaches are not directly applicable to this space. For that, we will map the trajectory in $SE(3) \times \ldots \times SE(3)$ to its Lie algebra $se(3) \times \ldots \times se(3)$, the tangent space at the identity element I_4. With this method, we will map all the trajectories from the Lie group to the same tangent space to the identity, and we argue that the mapped curves are quite faithful to the original curves because they are close to the identity element of the Lie group as they represent the transformations of the same body parts across successive frames. The resulting curve in the Lie algebra corresponding to $\delta(t)$ is given by:

$$\begin{aligned} \sigma(t) = (vec(log(P_{f_t(1),f_{t+1}(1)}(t))), vec(log(P_{f_t(2),f_{t+1}(2)}(t))) \\ \ldots, vec(log(P_{f_t(M),f_{t+1}(M)}(t)))) \in se(3) \times \ldots \times se(3) \end{aligned} \tag{5}$$

The dimension of the characteristic vector at any time t of $\sigma(t)$ is equal to $6M$. For this, the temporal representation of the action sequence is a vector of dimension $6 \times M \times N$, where M is the number of parts ($M = 19$ parts), and N represents the number of frames in the sequence.

3.3. Object Modeling

3.3.1. Object Detection

In order to deal with human–object interactions, one key step is to recognize the object from an RGB image. For this, we propose to use an object recognition algorithm based on the neural network [37]. The you only look once (YOLO) model processes images in real time at 45 frames per second. A smaller version of the network, Fast YOLO, processes 155 frames per second while still achieving double the mAPof other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors, but is less likely to predict false positives on the background. YOLO system detection is a regression problem. It divides the image into an even grid S × S and simultaneously predicts bounding boxes B, confidence in those boxes, and class probabilities C. These predictions are encoded as an S × S × (B × 5 + C) tensor. YOLO imposes strong spatial constraints on the box prediction delineation since each cell in the grid predicts only two boxes and can only have one class. This spatial constraint limits the number of nearby objects that the model can predict. Figure 3 shows the steps of detecting objects in an RGB image.

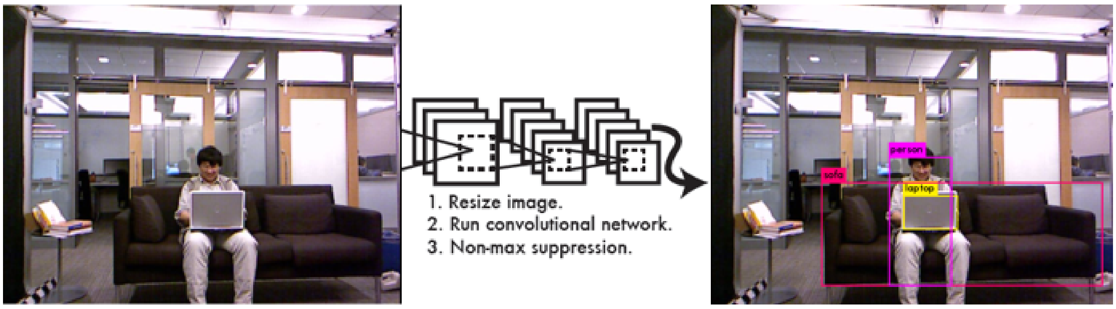


Figure 3. The YOLO detection system. (1) resizes the input image to 448 × 448; (2) runs a single convolutional network on the image; and (3) thresholds the resulting detections by the model's confidence.

3.3.2. Object Trajectory

Once having been detected in 2D, the object is tracked in 3D using the ICP algorithm. The resulting successive transformations are modeled as a trajectory in $SE(3)$. This trajectory is then mapped to the Lie algebra $se(3)$ and is fused with the trajectory in $se(3) \times \ldots \times se(3)$ $(n-1$ times) generated by the body parts. The trajectories modeling the activity lie in $se(3) \times \ldots \times se(3)$ $(n$ times) and have to be compared independently of the execution rate. Therefore, they are considered as time parameterized curves, and an elastic metric will be used to provide a time re-parameterization-invariant metric. The additional trajectory (generated by the object) is used only when comparing the proposed approach to RGB-D-based approaches. In this case, the output of the last layer of the object detection neural network applied on the first frame is also used (when the color channel is considered) to build the final feature vector.

4. Rate Invariance Modeling and Classification

We start by outlining a mathematical framework for helping in analyzing the temporal evolution of human activity when viewed as trajectories on the shape space of parametrized curves. This framework respects the underlying geometry of the shape of the trajectories, seen as curves, and helps maintain the desired invariance, especially re-parameterization of the trajectory curve that represents the execution rate. The next step is to calculate the distance between a given trajectory (to classify) to all training ones; let k trajectories be in the training set, resulting in a k-dimensional feature vector.

4.1. Elastic Metric for Trajectories

This representation has been used previously in biometric and soft-biometric application [38–43]. In our case, we will analyze the shape of the trajectories by the square root velocity function (SRVF) $q : I \rightarrow \mathbb{R}^n$ defined as:

$$q(t) = \frac{\sigma(t)}{\sqrt{||\sigma(t)||}} \tag{6}$$

$q(t)$ is a special function introduced in [35] that captures the form of $\sigma(t)$ while offering easy calculations, and the L^2 norm represents the metric that allows us to compare the shape of two trajectories. The set of all trajectories, denoted as C, is thus defined as follows:

$$C = \{q : I \rightarrow \mathbb{R}^n | ||q|| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^n) \tag{7}$$

$||.||$ is the norm. With the norm on its tangent space, C becomes a Riemannian manifold named the pre-shape space. Each element represents a trajectory in $\mathbb{R}^n$. We define the distance between two elements q_1 and q_2 by the length of the geodesic path between q_1 and q_2 on the variety C. The geodesic path between any two points $q_1, q_2 \in \mathcal{C}$ is given by the great circle, $\psi : [0, 1] \rightarrow \mathcal{C}$, where:

$$\psi(\tau) = \frac{1}{\sin(\theta)} \times (\sin((1-\tau)\theta) \times q_1 + \sin(\theta\tau) \times q_2) \tag{8}$$

The geodesic length is $\theta = d_{\mathcal{C}}(q_1, q_2) = cos^{-1}(< q_1, q_2 >)$. Let us define the equivalent class of q as: $[q] = \{\sqrt{\dot{\gamma}(t)} \times q(\gamma(t)),\ \ \gamma \in \Gamma\}$. The set of such equivalence classes, denoted by $\mathcal{S} \doteq \{[q] | q \in \mathcal{C}\}$, is called the shape space of open curves in $\mathbb{R}^n$. As shown in [35], $\mathcal{S}$ inherits a Riemannian metric from the larger space $\mathcal{C}$ due to the quotient structure. To obtain geodesics and geodesic distances between elements of $\mathcal{S}$, one needs to solve the optimization problem:

$$\gamma^* = argmin_{\gamma \in \Gamma} d_c(q_1, \sqrt{\dot{\gamma}} \times (q_2 \circ \gamma)). \tag{9}$$

The optimization over Γ is done using the dynamic programming algorithm. Let $q_2^*(t) = \sqrt{\gamma^* \dot{(t)}} \times q_2(\gamma^*(t)))$ be the optimal element of $[q_2]$, associated with the optimal re-parameterization γ^* of the second curve, then the geodesic distance between $[q_1]$ and $[q_2]$ in $\mathcal{S}$ is $d_s([q_1],[q_2]) \doteq d_c(q_1, q_2^*)$, and the geodesic is given by Equation (8), with q_2 replaced by q_2^*.

4.2. Feature Vector Building and Classification

We propose four variants of our method based on the channels used. The first one (geometric G) uses only the skeleton data. The second one (G + D) uses the skeleton and the depth channel. The third variant (G + C) uses the skeleton and the color channels. The last variant uses all channels (G + D + C). We present first the feature vector for the geometric G approach. Let n be the number of joints in a skeleton, the spatio-temporal modeling presented in Section 3, and k the number of trajectories in the training set with labels $l_1, \ldots, l_k$. For a given sequence in the test set, the first step is to represent it as a trajectory in $SE(3)^{n-1}$ as described in Section 3. Then, the elastic framework is applied in order to compute the elastic distance from the given trajectory to each of the k training ones. As illustrated in the previous section, the use of the elastic metric in trajectories' comparison ensures a rate-invariant distance. The resulting vector of k distances represents the feature vector of the geometric approach (G). An additional trajectory in $SE(3)$ must be considered when the depth data are used in the (G + D) approach. The feature vector has the same size; however, the trajectories are considered in $SE(3)^n$ rather than $SE(3)^{n-1}$. When the color channel is considered, the output of the last layer in the deep network used for object detection is concatenated to the k distance in order to build the feature vector denoted by $FeatureV$. The steps of feature vector building and classification are illustrated in Algorithm 1.

The resulting feature vector is fed to the Hoeffding tree (VFDT) algorithm. The Hoeffding tree [36] or very fast decision tree (VFDT) is built incrementally over time by splitting nodes (into two) using a small amount of the incoming data stream. The number of samples considered by the learning to expand a node depends on a statistical method called the Hoeffding bound or additive Chernoff bound. The Hoeffding tree is constructed by making recursive splits of leaves from a blank root and subsequently getting internal decision nodes, such that a tree structure is formed. The splits are decided by heuristic evaluation functions that evaluate the merit of the split-test based on attribute values.

Algorithm 1 Action sequences' classification.

```
Input:
k+1 sequences: $S_1, \ldots, S_k$ k training sequences of size $w_1, \ldots, w_k$, respectively, and $S_{k+1}$ test sequence of size $w_{k+1}$.
r: number of body parts (r = n − 1).
Output:
Label of sequence $S_{k+1}$
Begin
for $i = 1$ to $k + 1$ do
    OBJ-RGB(i)(1) = YOLO($S_i$(1))
    $OBJ - Depth(i)(1) = RGB2Depth(OBJ - RGB(i)(1))$
    for $j = 1$ to $w_i - 1$ do
        ($R_{i,j}, d_{i,j}$) = ICP (OBJ-Depth(i)(j), OBJ-Depth(i)(j+1))
        for $l = 1$ to $r$ do
            $P_{i,j,l} = (R_{i,j,l}, d_{i,j,l}) = procruste(part_{i,j,l}, part_{i,j+1,l})$
            $\sigma(i,j,l) = vec(log_{SE(3)}(P_{i,j,l}))$
        end for
    end for
end for
$q_{k+1} = \frac{\sigma(k+1)}{\sqrt{\|(\sigma(k+1,)))\|}}$
for $i = 1$ to $k$ do
    $q_i = \frac{\sigma(i)}{\sqrt{\|(\sigma(i)))\|}}$
    $\gamma^* = argmin_{\gamma \in \Gamma} d_c(q_1, \sqrt{\dot{\gamma}}(q_2 \circ \gamma)).$
    $q_i^* = \sqrt{\dot{\gamma}^*}.q_i(\gamma^*))$
    $d_i = d(S_i, S_k + 1) = cos^{-1}(< q_i^*, q_{k+1} >)$
end for
FeatureV = concatenate($d_1, .., d_k, OBJ - RGB(k + 1)(1)$)
label ($S_k + 1$) = VFDT (FeatureV)
End
```

5. Experimentation and Results

In order to validate our method, an evaluation was conducted on three databases that represent different challenges, namely Microsoft Research (MSR) Action3D dataset [8], MSR-Daily Activity 3D [23], and the SYSU 3D Human-Object Interaction Set [44].

5.1. MSR Action 3D

5.1.1. Data Description and Protocol

The MSR-Action 3D dataset is a set of RGBD data captured by a Kinect. This dataset includes 20 actions performed by 10 different subjects facing the camera. Each action is performed two or three times, resulting in a total of 557 action sequences. 3D joint positions are extracted from the depth sequence using the real-time skeleton tracking algorithm proposed in [45]. All actions are performed without interaction with the objects. Two main challenges are identified: the strong similarity between the different groups of actions and the changes in the speed of execution of the actions. For each sequence, the dataset provides information about depth, color, and skeleton. As indicated in [8], ten sequences are not used in the experiments because the skeletons are missing or too erroneous. For our experiments, we use 547 sequences. In this dataset, we followed the same protocol of the cross topic of [8], in which half of the subjects are used for training and the other half for testing. Subjects 1, 3, 5, 7, and 9 are used for training and Subjects 2, 4, 6, 8, and 10 for testing. In [8], the sequences were divided into three subsets AS_1, AS_2, and AS_3, each containing eight actions. Sets AS_1 and AS_2 are intended to group actions with similar movements, while AS_3 is intended to group complex actions.

5.1.2. Experimental Result and Comparison

No action, in this dataset, includes an interaction with the object. Thus, the skeleton-based approach (G) is performed. Table 1 reports the recognition performance on MSR-Action3D compared to several state-of-the-art approaches: joint positions (JPs) [14]: concatenation of the 3D coordinates of all the joints $v_1,\ldots,v_N$; pairwise relative positions of the joints (RJPs) [16]: concatenation of all the vectors; joint angles (JAs) [5]: concatenation of the quaternions corresponding to all joint angles (we also tried Euler angles and Euler axis-angle representations for the joint angles, but quaternions gave the best results); individual body part locations (BPLs) [46]: each individual body part is represented as a point in $SE(3)$ using its rotation and translation relative to the global x-axis.

In the last row of Table 1, the average recognition rate for the three subsets AS_1, AS_2, and AS_3 is reported. The recognition rates of our approach on AS_1, AS_2, and AS_3 were 94.66%, 85.08%, and 96.76%, respectively. The accuracy on subset AS_2 was lower than the two other subsets. This behavior is similar to the state-of-the-art approaches as revealed in Table 1. The average accuracy of the proposed representations was 92.16%, which is superior to the performance of previous state-of-the-art approaches provided in Table 1.

Table 2 compares the proposed approach with various approaches to recognizing human actions on skeletons using the protocol of [8]. Here, we see that our approach is competitive with the state-of-the-art with a recognition rate equal to 92.16%.

Table 1. Recognition performance on the MSR-Action3D for different feature spaces using the protocol of [8]. JP, joint position; RJP, relative position of the joint; JA, joint angle; BPL, body part location.

Dataset	JP [14]	RJP [16]	JA [5]	BPL [46]	Proposed
AS_1	91.65	92.15	85.80	83.87	**94.66**
AS_2	75.36	79.24	65.47	75.23	**85.08**
AS_3	94.64	93.31	94.22	91.54	**96.76**
Average	87.22	88.23	81.83	83.54	**92.16**

Table 2. Comparison with the state-of-the-art results.

MSR-Action3D Dataset (Protocol of [8])	
Histograms of 3D joints [47]	78.97
EigenJoints [16]	82.30
Joint angle similarities [5]	83.53
Spatial and temporal part sets [48]	90.22
Co-variance descriptors [14]	90.53
Random forests [19]	90.90
Body parts (BP)+SRVF [20]	92.10
Intra-frame modeling [2]	92.49
Proposed approach: skeleton	**92.16**

5.2. MSR Daily Activity 3D

5.2.1. Data Description and Protocol

The MSR Daily Activity 3D dataset [23] is a set of RGB-D sequences of human sequences acquired with the Kinect. It contains 16 types of activities: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down. Each of them was performed twice by 10 subjects [23]. The dataset contains 320 videos = 16 × 10 × 2 (10 actors and two essays/actor). There are 20 body joints recorded, whose positions are quite noisy due to two poses: “sitting on sofa” and “standing close to sofa”.

The experimental protocol is the same as in [23], which divides the dataset into three subsets, AS_1, AS_2, and AS_3, as shown in Table 3.

Table 3. Subsets of actions, AS_1, AS_2, and AS_3 in the MSR Daily Activity 3D dataset [23].

AS_1	AS_2	AS_3
eat	drink	use laptop
read book	call cellphone	cheer up
write on a paper	use vacuum cleaner	play guitar
use laptop	sit still	stand up
toss paper	play game	sit down
walk	lie down on sofa	

5.2.2. Experimental Result and Comparison

Table 4 reports the results of our algorithm on the MSR Daily activity dataset. The average recognition rate using only the skeleton data is 87.55%. When the dynamics of the object is considered, we have an average recognition rate equal to 88%. Combining the feature vector resulting from the geometry of the skeleton and object to the appearance of the object yields good improvement of the recognition rate. Actually, using the geometry of the skeleton (G) and the appearance of the object (C), the average recognition rate is 94.44%. The performance is also improved by using in addition the geometry of the object (D) to reach a 95% recognition rate, which is very competitive compared with recent state-of-the-art approaches.

Table 4. Recognition performance on the MSR-DailyActivity3D dataset for different feature spaces: (D) depth; (C) color (or RGB); (G) geometry or skeleton.

Methods	**Accuracy %**
(G) Dynamic Temporal Warping [49]	54
(G) 3D Joints and Local occupancy patterns (LOP) [50]	78
(G) Histogram of Oriented 4D Normals (HON4D) [3]	80.00
(G) Spar-Sity learning to Fuse atomic Features (SSFF) [27]	81.9
(G) Deep Model-Restricted Graph-based Genetic Programming (RGGP) [26]	85.6
(G) Action-let Ensemble [50]	85.75
(G) Super Normal [10]	86.25
(G) Bilinear [51]	86.88
(G) Depth Cuboid Similarity Feature (DCSF) + Joint [52]	88.2
(G) Local Flux Feature (LFF) + Improved Fisher Vector (IFV) [28]	91.1
(G) Group Sparsity [12]	95
(G) Range Sample [11]	95.6
(G) Heterogeneous Feature Machines (HFM) [53]	84.38
(G) Model of Probabilistic Canonical Correlation Analyzers (MPCCA) [54]	90.62
(G) Multi-Task Discriminant Analysi (MTDA) [55]	90.62
(G + D + C) JOULE [44]	95
Our Method:(G) Skeleton	**87.55**
Our Method:(G + D) Skeleton + Obj(D)	**88**
Our Method:(G + C) Skeleton + Obj(RGB)	**94.44**
Our Method:(G + D + C) Skeleton + Obj(RGB) + Obj(D)	**95**

5.3. SYSU 3D Human-Object Interaction Set

5.3.1. Data Description and Protocol

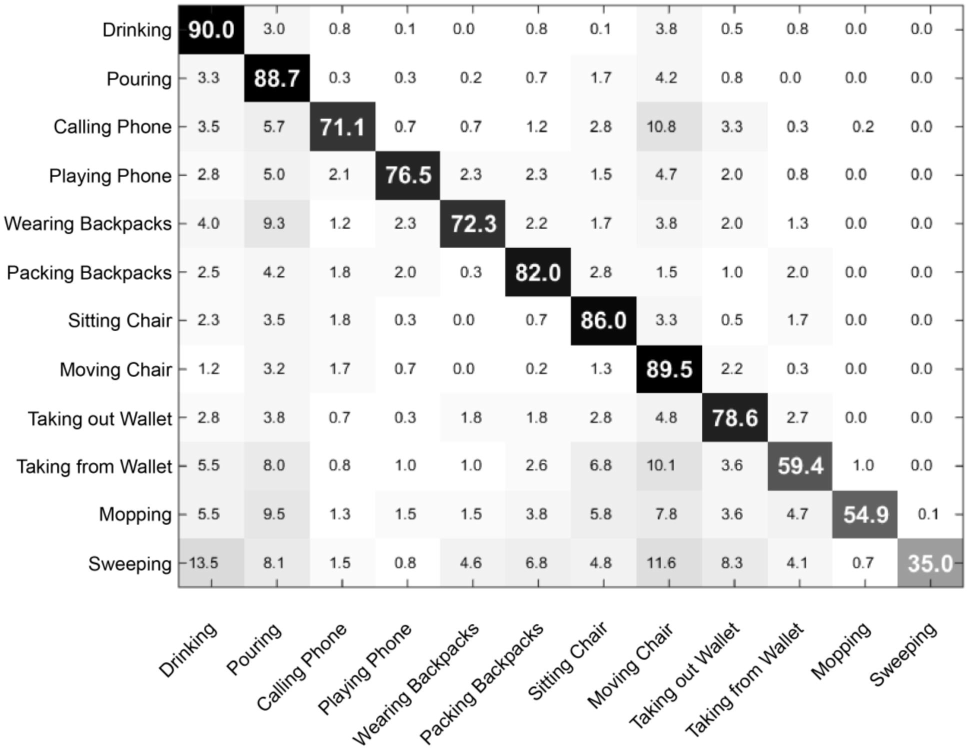
In this dataset [44], twelve different activities focusing on interactions with objects were performed by 40 persons. For each activity, each participant manipulates one of the six different objects: phone, chair, bag, wallet, mop, and besom. Therefore, there are in total 480 video clips collected in this set. For each video clip, the data acquisition is done by a Kinect camera, and we have the corresponding

RGB images, the depth sequence, and the skeleton. We tested all the methods compared with the second setting (Setting 2) [44]. The video footage made by half of the participants was used to learn the parameter model and the rest for the tests. We report the average precision and the standard deviation of the results on 30 random divisions For each parameter.

5.3.2. Experimental Result and Comparison

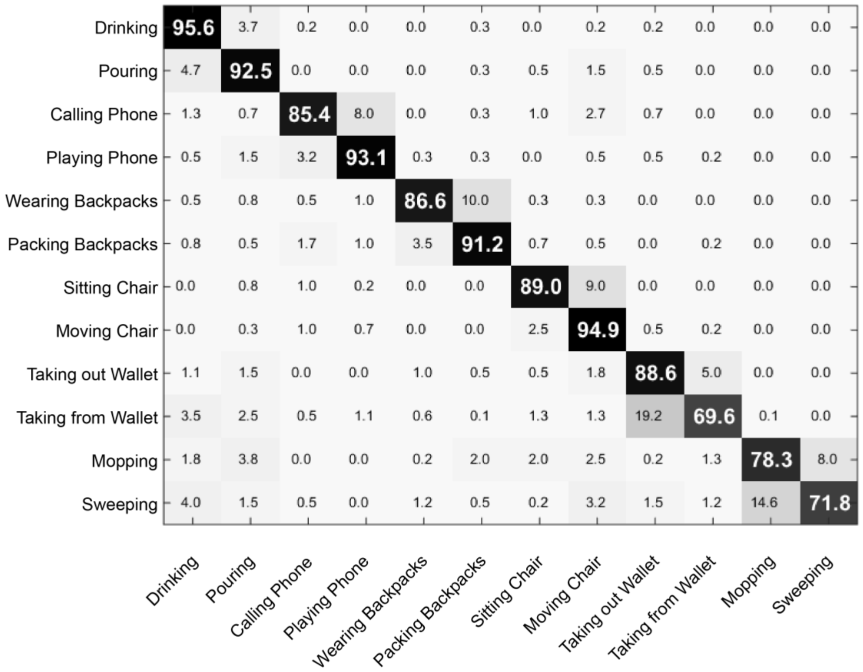
Table 5 provides the results of the different variants of the proposed approach compared to the state-of-the-art. When only the geometry of the skeleton data is considered, the recognition rate is 73.48%. This result is competitive compared to previous geometric approaches (based only on skeleton data). If we take into account the dynamics of the object, the recognition rate is equal to 74.51%. When the appearance of the object is considered in addition to the geometry of the skeleton, the recognition rate reaches 86.76 %, which represents the best recognition rate compared to the recent state-of-the-art approaches. The full version of the proposed approach, which makes use of all RGB-D and skeleton information, provides a recognition rate of 87.40%.

For further analysis of the obtained results, we illustrate in Figures 4 and 5 the confusion matrices on the SYSU dataset using the skeleton data (G) and the RGB-D (G + D + C) channels, respectively. The appearance of the object improves the performance of all actions, but the improvement is more considerable for sweeping and mopping actions: the skeleton data performs good for several actions, but seems not sufficient to distinguish other actions such as sweeping, with recognition rates of 35% and 54.9%, respectively. The skeleton motion during these two actions is similar to the motion while drinking, moving the chair, or pouring. The appearance of the object improves the performance for the sweeping and mopping actions. As shown in Figure 5, the recognition of sweeping improved by 36% by using the geometry and the appearance of the object in addition to the skeleton motion. The performance of mopping reaches 78.3% for the mopping action.



	Drinking	Pouring	Calling Phone	Playing Phone	Wearing Backpacks	Packing Backpacks	Sitting Chair	Moving Chair	Taking out Wallet	Taking from Wallet	Mopping	Sweeping
Drinking	90.0	3.0	0.8	0.1	0.0	0.8	0.1	3.8	0.5	0.8	0.0	0.0
Pouring	3.3	88.7	0.3	0.3	0.2	0.7	1.7	4.2	0.8	0.0	0.0	0.0
Calling Phone	3.5	5.7	71.1	0.7	0.7	1.2	2.8	10.8	3.3	0.3	0.2	0.0
Playing Phone	2.8	5.0	2.1	76.5	2.3	2.3	1.5	4.7	2.0	0.8	0.0	0.0
Wearing Backpacks	4.0	9.3	1.2	2.3	72.3	2.2	1.7	3.8	2.0	1.3	0.0	0.0
Packing Backpacks	2.5	4.2	1.8	2.0	0.3	82.0	2.8	1.5	1.0	2.0	0.0	0.0
Sitting Chair	2.3	3.5	1.8	0.3	0.0	0.7	86.0	3.3	0.5	1.7	0.0	0.0
Moving Chair	1.2	3.2	1.7	0.7	0.0	0.2	1.3	89.5	2.2	0.3	0.0	0.0
Taking out Wallet	2.8	3.8	0.7	0.3	1.8	1.8	2.8	4.8	78.6	2.7	0.0	0.0
Taking from Wallet	5.5	8.0	0.8	1.0	1.0	2.6	6.8	10.1	3.6	59.4	1.0	0.0
Mopping	5.5	9.5	1.3	1.5	1.5	3.8	5.8	7.8	3.6	4.7	54.9	0.1
Sweeping	13.5	8.1	1.5	0.8	4.6	6.8	4.8	11.6	8.3	4.1	0.7	35.0

Figure 4. SYSU 3D Human-Object Interaction dataset confusion matrix based on skeleton data (G).



	Drinking	Pouring	Calling Phone	Playing Phone	Wearing Backpacks	Packing Backpacks	Sitting Chair	Moving Chair	Taking out Wallet	Taking from Wallet	Mopping	Sweeping
Drinking	95.6	3.7	0.2	0.0	0.0	0.3	0.0	0.2	0.2	0.0	0.0	0.0
Pouring	4.7	92.5	0.0	0.0	0.0	0.3	0.5	1.5	0.5	0.0	0.0	0.0
Calling Phone	1.3	0.7	85.4	8.0	0.0	0.3	1.0	2.7	0.7	0.0	0.0	0.0
Playing Phone	0.5	1.5	3.2	93.1	0.3	0.3	0.0	0.5	0.5	0.2	0.0	0.0
Wearing Backpacks	0.5	0.8	0.5	1.0	86.6	10.0	0.3	0.3	0.0	0.0	0.0	0.0
Packing Backpacks	0.8	0.5	1.7	1.0	3.5	91.2	0.7	0.5	0.0	0.2	0.0	0.0
Sitting Chair	0.0	0.8	1.0	0.2	0.0	0.0	89.0	9.0	0.0	0.0	0.0	0.0
Moving Chair	0.0	0.3	1.0	0.7	0.0	0.0	2.5	94.9	0.5	0.2	0.0	0.0
Taking out Wallet	1.1	1.5	0.0	0.0	1.0	0.5	0.5	1.8	88.6	5.0	0.0	0.0
Taking from Wallet	3.5	2.5	0.5	1.1	0.6	0.1	1.3	1.3	19.2	69.6	0.1	0.0
Mopping	1.8	3.8	0.0	0.0	0.2	2.0	2.0	2.5	0.2	1.3	78.3	8.0
Sweeping	4.0	1.5	0.5	0.0	1.2	0.5	0.2	3.2	1.5	1.2	14.6	71.8

Figure 5. SYSU 3D Human-Object Interaction dataset confusion matrix based on skeleton, depth, and color data (G + D + C).

Table 5. Comparison on the SYSU 3D dataset. (D) depth; (C) color (or RGB); (G) geometry or skeleton.

Methods	Accuracy %
(G) Local Accumulative Frame Feature (LAFF) [56]	54.2
(G) Dynamic skeletons [44]	75.5 ± 3.08
(G) LSTM-trust gate [57]	76.5
(G + D + C) LAFF [56]	80
(G + D + C) JOULE [44]	84.9 ± 2.29
Our Method:(G) Skeleton	**73.48 ± 5.91**
Our Method:(G + D) Skeleton + Obj (D)	**74.51 ± 5.47**
Our Method:(G + C) Skeleton + Obj (RGB)	**86.76 ± 4.82**
Our Method:(G + D + C) Skeleton + Obj (RGB) + Obj (D)	**87.40 ± 5.04**

6. Conclusions and Future Direction

In this paper, we represent the inter-frame evolution of skeleton body parts in Lie group $SE(3) \times \ldots \times SE(3)$. When an object is involved in the action, a neural network is used to detect the object at the first frame, then the evolution across frames is tracked, then similarly modeled as an additional trajectory in Lie group $SE(3)$. The resulting trajectories are then mapped onto the Lie algebra, where they are compared using a re-parameterization-invariant framework in order to handle rate variations. The distances to training trajectories are concatenated with the output of the last layer of the neural network used for object detection, then are fed to the very fast decision tree to perform action recognition. We experimentally show that the proposed approach performs better than many previous approaches for human activity recognition. As future work, we expect widespread applicability in domains such as physical therapy and rehabilitation.

Author Contributions: Formal analysis, H.D. and I.R.F.; Methodology, M.B., H.D. and I.R.F.; Supervision, M.M. and I.R.F.; Validation, H.D., M.M. and I.R.F.; Visualization, I.R.F.; Writing–original draft, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; Moore, R. Real-time human pose recognition in parts from single depth images. *Commun. ACM* **2013**, *56*, 116–124. [CrossRef]
2. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
3. Oreifej, O.; Liu, Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 716–723.
4. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *J. Vis. Commun. Image Represent.* **2014**, *25*, 24–38. [CrossRef]
5. Ohn-bar, E.; Trivedi, M.M. Joint Angles Similarities and HOG for Action Recognition. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013.
6. Vemulapalli, R.; Chellapa, R. Rolling Rotations for Recognizing Human Actions From 3D Skeletal Data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4471–4479.
7. Boujebli, M.; Drira, H.; Mestiri, M.; Farah, I.R. Rate invariant action recognition in Lie algebra. In Proceedings of the 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fez, Morocco, 22–24 May 2017.
8. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3d points. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 9–14.
9. Zhu, Y.; Chen, W.; Guo, G. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image Vis. Comput.* **2014**, *32*, 453–464. [CrossRef]
10. Yang, X.; Tian, Y. Super normal vector for activity recognition using depth sequences. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 804–811.
11. Lu, C.; Jia, J.; Tang, C.-K. Range-sample depth feature for action recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 772–779.
12. Luo, J.; Wang, W.; Qi, H. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013.
13. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
14. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In Proceedings of the International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; pp. 2466–2472.
15. Lv, F.; Nevatia, R. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 359–372.
16. Yang, X.; Tian, Y. Eigenjoints-based action recognition using naivebayes-nearest-neighbor. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 14–19.
17. Lillo, I.; Soto, A.; Niebles, J.C. Discriminative hierarchical modeling of spatio-temporally composable human activities. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 812–819.
18. Zanfir, M.; Leordeanu, M.; Sminchisescu, C. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2752–2759.

19. Zhu, Y.; Chen, W.; Guo, G. Fusing spatio-temporal features and joints for 3d action recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 486–491.
20. Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; Del Bimbo, A. 3D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold. *IEEE Trans. Cybern.* **2015**, *45*, 1340–1352. [CrossRef] [PubMed]
21. Meng, M.; Drira, H.; Boonaert, J. Distances evolution analysis for online and off-line human–object interaction recognition. *Image Vision Comput.* **2018**, *70*, 32–45. [CrossRef]
22. Meng, M.; Drira, H.; Daoudi, M.; Boonaert, J. Human-object interaction recognition by learning the distances between the object and the skeleton joints. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015.
23. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In Proceedings of the IEEE International Conference, Providence, RI, USA, 16–21 June 2012.
24. Chaudhry, R.; Ofli, F.; Kurillo, G.; Bajcsy, R.; Vidal, R. Bio-inspired Dynamic 3D Discriminative Skeletal Features for Human Action Recognition. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013.
25. Guo, S.; Pan, H.; Tan, G.; Chen, L.; Gao, C. A High Invariance Motion Representation for Skeleton-Based Action Recognition. *Int. J. Pattern Recognit. Artif. Intell.* **2016**, *30*, 1650018. [CrossRef]
26. Liu, L.; Shao, L. Learning discriminative representations from rgb-d video data. In Proceedings of the International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; pp. 1493–1500.
27. Shahroudy, A.; Wang, G.; Ng, T.-T. Multi-modal feature fusion for action recognition in rgb-d sequences. In Proceedings of the International Symposium on Control, Communications, and Signal Processing, Athens, Greece, 21–23 May 2014; pp. 1–4.
28. Yu, M.; Liu, L.; Shao, L. Structure-preserving binary representations for rgb-d action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1651–1664. [CrossRef] [PubMed]
29. Chaaraoui, A.A.; Padilla-Lopez, J.R.; Florez-Revuelta, F. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 1–8 December 2013; pp. 91–97.
30. Koppula, H.S.; Gupta, R.; Saxena, A. Learning human activities and object affordances from rgb-d videos. *Int. J. Robot. Res.* **2013**, *32*, 951–970. [CrossRef]
31. Lei, J.; Ren, X.; Fox, D. Fine-grained kitchen activity recognition using rgb-d. In Proceedings of the ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 208–211.
32. Wen, Z.; Yin, W. A feasible method for optimization with orthogonality constraints. *Math. Program.* **2013**, *142*, 397–434. [CrossRef]
33. Zhao, Y.; Liu, Z.; Yang, L.; Cheng, H. Combing rgb and depth map features for human activity recognition. In Proceedings of the IEEE Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Hollywood, CA, USA, 3–6 December 2012; pp. 1–4.
34. Murray, R.M.; Li, Z.; Sastry, S.S. *A Mathematical Introduction to Robotic Manipulation*; CRC Press: Boca Raton, FL, USA, 1994.
35. Joshi, S.H.; Klassen, E.; Srivastava, A.; Jermyn, I. A novel representation for riemannian analysis of elastic curves in Rn. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–7.
36. Xu, W.; Qin, Z. Constructing Decision Trees for Mining High-speed Data Streams. *Chin. J. Electron.* **2012**, *21*, 215–220.
37. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
38. Drira, H.; Ben Amor, B.; Srivastava, A.; Daoudi, M.; Slama, R. 3D Face Recognition under Expressions, Occlusions, and Pose Variations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2270–2283. [CrossRef] [PubMed]
39. Xia, B.; Ben Amor, B.; Drira, H.; Daoudi, M.; Ballihi, L. Combining face averageness and symmetry for 3D-based gender classification. *Pattern Recognit.* **2015**, *48*, 746–758. [CrossRef]

40. Xia, B.; Amor, B.B.; Drira, H.; Daoudi, M.; Ballihi, L. Gender and 3D facial symmetry: What's the relationship? In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013.
41. Amor, B.B.; Drira, H.; Ballihi, L.; Srivastava, A.; Daoudi, M. An experimental illustration of 3D facial shape analysis under facial expressions. *Ann. Telecommun.* **2009**, *64*, 369–379 [CrossRef]
42. Mokni, R.; Drira, H.; Kherallah, M. Combining shape analysis and texture pattern for palmprint identification. *Multimed. Tools Appl.* **2017**, *76*, 23981–24008 [CrossRef]
43. Xia, B.; Amor, B.B.; Daoudi, M.; Drira, H. Can 3D Shape of the Face Reveal your Age? In Proceedings of the 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, 5–8 January 2014; pp. 5–13.
44. Hu, J.-F.; Zheng, W.-S.; Lai, J.; Zhang, J. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2186–2200. [CrossRef] [PubMed]
45. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp,T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the CVPR, Providence, RI, USA, 20–25 June 2011.
46. Yacoob, Y.; Black, M.J. Parameterized Modeling and Recognition of Activites. In Proceedings of the Sixth International Conference on Computer Vision, Bombay, India, 4–7 January 1998.
47. Xia, L.; Chen, C.C.; Aggarwal, J.K. View Invariant Human Action Recognition Using Histograms of 3D Joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012.
48. Wang, C.; Wang, Y.; Yuille, A.L. An Approach to Pose-based Action Recognition. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
49. Muller, M.; Roder, T. Motion templates for automatic classification and retrieval of motion capture data. In Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Vienna, Austria, 2–4 September 2006; pp. 137–146.
50. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Learning actionlet ensemble for 3d human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 914–927. [CrossRef] [PubMed]
51. Kong, Y.; Fu, Y. Bilinear heterogeneous information machine for rgbd action recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1054–1062.
52. Xia, L.; Aggarwal, J. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2834–2841.
53. Cao, L.; Luo, J.; Liang, F.; Huang, T.S. Heterogeneous feature machines for visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1095–1102.
54. Cai, Z.; Wang, L.; Qiao, X.P.Y. Multi-view super vector for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 596–603.
55. Zhang, Y.; Yeung, D.-Y. Multi-task learning in heterogeneous feature spaces. In Proceedings of the Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011.
56. Hu, J.-F.; Zheng, W.-S.; Ma, L.; Wang, G.; Lai, J. Real-time RGB-D activity prediction by soft regression. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 280–296.
57. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. *Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition*; Springer: Cham, Switzerland, 2016; pp. 816–833.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.