Spatiotemporal representation of 3D skeleton joints-based action recognition using modified spherical harmonics

Adnan Al Alwani, Youssef Chahir

Spatiotemporal Representation of 3D Skeleton Joints-Based Action Recognition using Modified Spherical Harmonics

Adnan Salih AL ALWANI[a], Youssef CHAHIR[a]

[a]GREYC CNRS (UMR 6072), University of Caen Basse-normandie, Caen 14032, France

ABSTRACT

Action recognition based on the 3D coordinates of body skeleton joints is an important topic in computer vision applications and humanrobot interaction. At present, most 3D data are captured using recently introduced economical depth sensors. In this study, we explore a new method for skeleton-based human action recognition. In this novel framework, the normalized angles of local joints are first extracted, and then the modified spherical harmonics (MSHs) are used to explicitly model the angular skeleton by projecting the spherical angles onto the unit sphere basis. This process decomposes the skeleton representation into a set of basis functions. A spatiotemporal system of the spherical angles is adopted to construct the static pose and joint displacement over a human action sequence. Consequently, the MSH coefficients of the joints are used as the discriminative descriptor of the sequence. The extreme learning machine (ELM) classifier and recently published 3D action datasets are used to validate the proposed method. The experimental results show that the proposed approach performs better than many classical methods.

1. Introduction

Action representation and recognition from 3D data are among the topics widely discussed in pattern recognition. In recent years, research trends have mainly shifted toward the action recognition of sequences captured by RGB-D cameras. These ranging cameras feature the facility of 3D imaging technologies to provide 2D images and depth maps, and simplify the acquisition of 3D human posture through skeleton joints (Shotton et al.) . Moreover, the popularity of low-cost depth sensors, such as the Kinect sensor, has led to the development of efficient methods for specific action recognition applications (Raptis et al., 2011; Wang et al., 2012). The studies recently conducted by (Alnowami et al., 2012; Obdrzalek et al., 2012; Yao et al., 2011) indicate that, compared with traditional 2D image data, the representation of skeleton joints only for action recognition provides better results. Therefore, the current study uses skeleton joint data as an initial input for skeleton representation. With RGB-D stationary sensors, the sequences of depth maps can be registered as a set of skeleton joints in 3D coordinates (Shotton et al.). In other words, the application of RGB-sensors allows human action recognition problems to be handled directly as 3D positions of joints, which cannot be realized with a silhouette-based imagery system. Importantly, depth sensors not only achieve fast capturing speed and good view invariance, but they also provide accurate depth maps (Li et al., 2010; Xia et al., 2012). Skeleton-based action recognition methods typically rely on the direct use of either the absolute position of skeleton joints or the connection between these joints (rigid segments between joints). Joint position approaches consider the human skeleton simply as a set of articulated points, in which only the skeleton joints are used to abstract body motion in 3D space. Likewise, these approaches investigate human motion either by individual joints or are relational between joints using various features, including joint positions (Hussein et al., 2013; Raptis et al., 2008), joint orientations with respect to a reference coordinate axis (Xia et al., 2012) and pairwise relative joint positions (Wang and Lee, 2009; Yang and Tian, 2012).

With the efforts of numerous researchers to reliably improve skeleton-based recognition algorithms, research attention

Corresponding author: Tel.: +0-33-0650871066,
Corresponding author:
e-mail: adnan.alalwani@unicaen.fr (Youssef CHAHIR), youssef.chahir@unicaen.fr (Youssef CHAHIR)

is now directed toward using skeleton joint data to recognize human actions. Loss of accuracy and effective feature representation remains a major problem in the field. This issue should be addressed to robustly handle the task of human action recognition by either improving the discriminative feature descriptions or pose representation or by relying on additional post-processing of skeletal data. Several difficulties remain in undertaking recognition tasks based on 3D joint representation (e.g., occlusion of body parts, lack of precision, and errors in data acquisition).

In this study, we present a novel skeleton joint-based representation of 3D human action in a spatiotemporal manner. We employ the spherical angles of body joints computed from the 3D coordinates of skeleton joints. The proposed feature representation is a combination of the modified spherical harmonics (MSHs) and the spatiotemporal model of sequence level. To estimate the human pose, the spherical harmonics (SHs) of spherical angles provide a distinctive feature description. As such, the problem of skeleton joint representation is addressed in a spatiotemporal approach using MSHs. The proposed model simply incorporates two mechanisms to efficiently capture the temporal dynamic of joints, namely, the application of MSHs in the computed spherical angles of each pose and the construction of MSHs in a hierarchical scheme. MSHs are computed in multi-levels, in which each level encodes the time window of an action sequence.

In the proposed representation of 3D human action, the selected MSHs are adopted to characterize the features in multi-levels and capture the harmonic frequency of function in a two-sphere space S^2. Given this condition, the defined spherical angle vector of the selected joints may be projected onto (S^2). However, the principle computation required in this space is extremely large because each selected joint is sampled by the feature vectors of $M_J = M_1, , M_k$, $M \in R^{N*N}$; where M is an MSH matrix of k levels, J joint numbers, and N number of frames in each level. Considering that the desired descriptor dimensionality aims to expedite the classification phase as well as reduce the noise and redundant feature sizes, we apply dynamic time wrapping (DTW) to determine the optimal alignment between the sub-levels of hierarchical MSHs.

Unlike recent works that rely directly on individual joint locations, our study is related to explicit skeleton model approaches. In particular, this study capitalizes on a new feature space that has not been previously considered. Therefore, all the human skeleton joints are represented as a collection of measured features in static pose and joint motion nature. An action classification is performed using the extreme learning machine (ELM) classifier and SH-based skeleton representation. The proposed method is evaluated based on recent skeleton-based 3D action datasets.

In this paper, we present an improved skeleton representation by applying the MSHs and spatiotemporal modeling of skeleton joints along the sequence level. The ELM is also compared with other classification algorithms using the same datasets.

2. Related Work

In this section, we briefly summarize various skeleton-based human action recognition approaches. In particular, the methods related to our developed technique are reviewed because they rely on skeleton joint data only. For a recent detailed survey on human motion analysis from depth data, see references (Aggarwal and Xia, 2014), and (Ye et al., 2013). 3D pose-based approaches have been explored by various researchers. (Yao et al., 2011) indicated that the application of skeleton data (e.g., positions, velocities and angles of a joint from a human articulated body) outperforms gray-based features captured by 2D cameras in an indoor environment scenario. In general, many useful features can be initially extracted from RGB-D skeletal data. The majority of these features can be divided into two: those that are based on the angular characteristics of joints and those that are based on the generic 3D coordinates of joints. In certain action recognition methods, the features are developed in complex models to form the representation of the motion sequences.

3D joint positions are commonly extracted as features through four mechanisms. First, raw 3D data are recognized directly without any further processing (Raptis et al., 2008; Shimada and Taniguchi, 2008; Wang and Lee, 2009). Second, these data are further processed to address certain challenges (Barnachon et al., 2013; Wang et al., 2012; Zhao et al., 2013). Third, the distances between each joint can be used as a distance-based feature vector for each frame (Antnio et al., 2012). Fourth, the features for the selected joints can be simply calculated with reference to the relative distance between joints (Wang et al., 2012).

In (Hussein et al., 2013), the human body skeleton was interpreted by directly constructing 3D skeleton joint locations as covariance descriptors, and the temporal evolutions of the action dynamic were modeled using a temporal hierarchy of covariance descriptors. In (Lv and Nevatia, 2006), the 3D coordinates of the joints were used for a skeleton representation of the human body. Correspondingly, the temporal nature of the action sequence was modeled with a generative discrete hidden Markov model (HMM), and action recognition was performed using the multiclass AdaBoost. The view-invariant representation of the human skeleton was proposed in (Xia et al., 2012) by partitioning the 3D spherical coordinates into angular spaced bins based on the aligned orientations with respect to a coordinate system registered at the hip center. A generative HMM classifier, which addresses the temporal nature of pose observations, was then used to classify each visual code word identified with the cluster method.

The proposed work in (Wang et al., 2012) applied the idea of the pairwise relative locations of joints to represent the human skeleton. The temporal displacement of this representation was characterized using the coefficients of a Fourier pyramid hierarchy. Moreover, the researchers proposed an action let-based approach, in which the effective joint combinations were selected using a multiple kernel learning approach. In (Yang and Tian, 2012), the skeleton joints were represented by combining the temporal and spatial joint relations. To explicitly model the motion displacement, the researchers adopted a method for

skeleton representation on the basis of relative joint positions, temporal motion of joints and offset of joints with respect to the reference frame. The resulting descriptors were projected onto eigenvectors using Principle Component Analysis PCA . In this case, each frame was described by an EigenJoint descriptor, and action recognition was performed using the nave Bayes nearest neighbor. The same scheme was used for the skeleton representation in (Zhu et al., 2013), in which action recognition was achieved by adopting the random forest classifier. The view-invariant action representation framework was proposed by (Evangelidis et al., 2014). In this work, the quad-based skeletal feature was adopted to encode the local relation between joints in quadruple form. Consequently, the 3D similarity invariance was achieved. The researchers also adopted a Fisher kernel representation based on a Gaussian mixture model. Such a representation generates the skeletal quads and invokes a multilevel splitting of sequences into segments to integrate the order of subactions into the vector representation. In (Raviteja et al.), a human skeleton was presented as points in the Lie group. The proposed representation explicitly models the 3D geometric relationships among various body parts using rotations and translations. Given that the Lie group was a curved manifold, the researchers mapped all action curves from the Lie group to its Lie algebra, and the temporal evolutions were modeled with DTW.

Our proposed method is different in many aspects over the previously published methods. To more explicitly examine these issue, we now include the comparison in terms of number of joints used, choice of classifier, computation complexity and dataset variability.

In term of number of joints, several techniques such as [(Yang and Tian, 2012; Hussein et al., 2013; Raviteja et al.; Evangelidis et al., 2014; Ohn Bar and Trivedi, 2013; Wang and Lee, 2009)] adopted all skeleton joints for feature extraction. Whereas, our method adopted only nine joints. For the classification task, we used ELM which provides a very high recognition accuracy with very fast training time compared to many other classification methods such as :

- (SVM) used (Hussein et al., 2013),
- (NBNN) used in (Yang and Tian, 2012) and (Seidenari et al., 2013).
- (AdaBoost) used in (Bloom et al., 2012),
- (HHM) used in (Xia et al., 2012).

The calculation of our features are computationally light than many previous works which try to classify a large 3D point, such in (Raptis et al., 2011), (Raviteja et al.), (Ohn Bar and Trivedi, 2013), (Zhu et al., 2013), and (Li et al., 2010). Moreover, in our method, we perform the evaluation on a variety of datasets compared to many approaches ((Xia et al., 2012), (Yang and Tian, 2012), and (Seidenari et al., 2013)).

The main contribution of our work was to reveal the potential of features that can be extracted from the 3D joint locations, without requiring the additional processing of the entire depth maps of a sequence, as in (Yang and Tian, 2012), (Zhu et al., 2013), and (Ohn Bar and Trivedi, 2013).

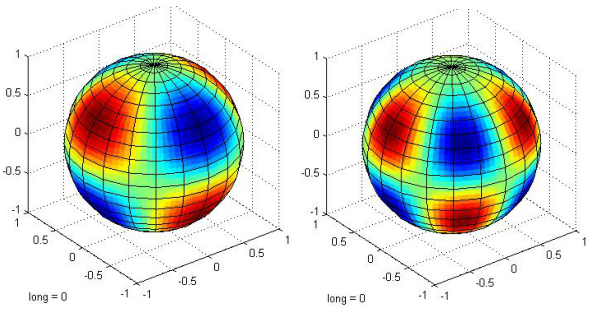

Fig. 1. Plots of the spherical harmonic basis functions. Blue indicates positive values and red indicates negative values.(Top), SHs for frequency l=3, and order m=2.(Bottom), frequency l=4, and order m=3.

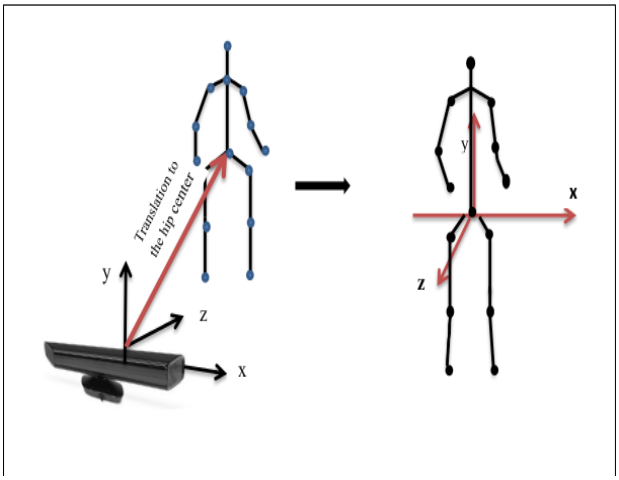

Fig. 2. 3-dimensional coordinates corresponding to a human body skeleton. (left) real world coordinate , (right) body reference coordinate transformation.

3. Methodology

3.1. Overview of Spherical Harmonics

SHs are Fourier series defined on the basis of a 2-sphere. Given that a Fourier series is a set of mathematical tools for expanding trigonometric functions, SHs are used to organize unit-sphere functions by angular frequency in terms of spherical coordinates. An illustration of a basis functions by means of their harmonics is configured into a set of rows as shown in Fig. 1. The colors in the Fig. 1 represent positive and negative values of SHs.

SHs are adopted in various fields to solve specific types of differential equations, such as the representation of gravitational fields and the modeling of geoscience computational problems. SHs have also been adopted to solve various problems in the context of computer vision applications. Such as, face recognition under unknown lighting challenge was adopted by (Zhang and Samaras, 2006), as well as 3D object retrieval (Bustos et al., 2005), and volumetric descriptors (Vranic, 2003).

In this section, we briefly explain SH theory. For a general introduction to SH transform, see (Freeden and Michael, 2009;

Silverman, 1972), which present the classical material on SHs. Let $(r, \theta, \phi) : r \in R^+, \theta \in [0, 2\pi], \phi \in [0, \pi]$ be the spherical coordinates, and $f(\theta, \phi)$ be the homogeneous harmonic functions on R^3. In this study, we aim to determine the homogeneous solutions of Laplaces equation, $\nabla^2 f = 0$, in spherical coordinates. Likewise, we intend to explain how these solutions correspond to the decomposition of eigenfunctions in space $L^2(S^2)$, $S^2 = (\theta, \phi, r) \in R^3$. In this case, SH is the generalization of Fourier series to 2-sphere by projecting the square-integrable function S^2 onto Hilbert space $L^2(S^2)$. However, for the spherical coordinates:

$$\begin{aligned} x &= r \sin\theta \cos\phi \\ y &= r \sin\theta \sin\phi \\ z &= r \cos\theta. \end{aligned} \tag{1}$$

where the Laplacian of a harmonic function on two-sphere using the spherical coordinates is given by

$$\mathbf{\Delta}_{\mathbf{S}^2} f = \frac{1}{\sin\theta} \frac{\partial}{\partial\theta} \left(\sin\theta \frac{\partial f}{\partial\theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2 f}{\partial\phi^2}. \tag{2}$$

The final solution of Laplacian in R^3 (the detailed solution can be found in (Silverman, 1972)) is a set of Legendre function and eigenfunctions expressed as follows:

$$f(\theta, \phi) = Q(Z_n^m(cos\theta))(\exp(jm\phi)). \tag{3}$$

where Q is a constant. The first term in Eq. 3 is a set of Legendre polynomials, and the second term is the eigenfunctions of the Laplacian on a sphere with an eigenvalue of $n(n + 1)$. The notation of the preceding equation represents the SHs in complex form. In this context, we adopt the notion of real SHs with the degree of n and order of $m > 0$. Thus, we set

$$y_n^m(\theta, \phi) = \sqrt{2} Q_n^m \cos(m\phi) Z_n^m(\cos\theta). \tag{4}$$

Q_n^m are the scaling factors expressed as

$$Q_n^m = \sqrt{\frac{(2n + 1)(n - |m|)!}{4m(n + |m|)!}}. \tag{5}$$

(Silverman, 1972) specified that any function of the form $f(\theta, \phi)$ can be represented by a set of expansion coefficients on the unit sphere. The complete harmonic basis functions are indexed by two integer constants (i.e., the degree n and the order m). The sampling frequencies of the basis functions over the unit sphere are defined by the values of the order $-n \leq m \leq n$. In general, there are $2n + 1$ bases. As an illustration example, the visual representations of the real SHs for the azimuth and elevation directions are shown in Fig. 1. The blue portions represent the positive harmonic functions, while the red portions depict the negative ones. The distance of the surface from the origin (rows) indicates the value of harmonics in angular direction (θ, ϕ).

3.1.1. Modified spherical harmonics (MSHs)

this section presents a proposed feature extraction framework, in which the modified real part notation of SHs is used to represent the spatiotemporal features of skeleton joints and improve human action recognition. The notation $\cos(m\phi)$ denotes the real part of SHs. For the special case of degree $m = 2$, the modified SHs can be computed given that $\cos(2\phi)$ is explicitly expressed as:

$$\cos(2\phi) = 2\cos^2\phi - 1. \tag{6}$$

Substituting Eq. 6 in Eq. 4 and rearranging the terms in the latter establish the modified real SHs as follows:

$$y_n^m = Q_n^m [2\cos^2\phi - 1] Z_n^m \cos\theta. \tag{7}$$

The quadratic term in 6 captures the angular velocity of joint displacement. This velocity is useful to differentiate the actions involved in a curved motion, such as waving or shape drawing. Thus, for a given action, the angular quantities (e.g., relative angular speed and changes in directions of these joints) can be more stable across objects than their actual 3D positions.
However, the MSHs of the local 3D skeleton joints capture discriminant information about different actions. In other words, the quadratic term in MSHs describes the direction and angular speed of joint motions. Experiments reported in section 6.2 have proven that introducing the quadratic angular velocity and direction of joint dynamics significantly improves the use of the standard SHs.

3.2. 3D Skeleton Coordinate

The raw data (3D coordinates of the joints) contain useful information about the motion sequence of a human. To estimate the skeleton features, a depth sensor is generally used to easily develop a skeleton model consisting of K joints. To make the 3D joint locations invariant to sensor parameters, the joint positions must be mapped into a unique coordinate system.
However, Since only the global 3D coordinates of the skeleton joints are available. Thus, to align the body skeleton with the reference coordinate system, we consider the origin of body coordinates indicated by the hip center and set the horizontal reference vector as the vector directed from the left hip to the right hip. Moreover, the 3D subject coordinates comprise three orthogonal vectors (γ, ρ, β) as depicted in Fig. 2. The first axis γ (horizontal axis) is from the right hip to the left one. The second axis (perpendicular axis) ρ is always directed toward the head and aligned with the vertical dimension of the torso. The second vector is computed by rotating γ vector by 90°. The third axis stems from the cross product of the two bases $(\beta = \gamma \times \rho)$. The selection of the $hip - centerand \quad Right/Lefthip$ joints ensures an acceptable approximation of the subject coordinates along the joint movement. At this preprocessing stage, a significant reduction of joints is usually performed because these joints are not involved in the final action representation.

3.3. Features Extraction

Body pose and local joint displacement must be incorporated into the 3D skeleton-based action descriptor. Therefore, we appropriately organize the skeleton joints in a manner where both

static pose and joint displacement can be reliably handled at a specific instance in time. In this context, the skeleton joints are represented in terms of the spherical angles relatively measured with respect to the fixed coordinates, which are more accurate than the joint coordinates or joint difference.
To characterize the body pose properties, the spherical angles are quantified in the spherical coordinate. All angles are computed corresponding to the origin reference (i.e., the origin of the spherical coordinate system is placed at the *hip − center* joint coordinate). Only a primitive set of the supported joints is used for the 3D pose representation. In this case, only the joints that correspond to *Right/Left elbow, Right/Left wrist, Right/Left knee, Right/Left foot, and Head* are selected. For every selected joint J_i, the following spherical angles are estimated over the action sequence:

$$\begin{aligned} \theta(t) &= \arctan(\tfrac{\gamma}{\rho}) \\ \phi(t) &= \arccos\left(\frac{\beta}{\sqrt{(\rho^2+\gamma^2+\beta^2)}}\right). \end{aligned} \tag{8}$$

where t is the frame index, and θ_i and ϕ_i are the estimated spherical angles.

3.4. Spatiotemporal-based skeleton joint representation

To further analyze the 3D skeleton joints in terms of their spatiotemporal domain, we propose a new joint representation using MSHs which it is an extension of SHs. SHs are a frequency domain basis for characterizing homogeneous functions defined over a two-sphere. They are the extension of the 1D Fourier series on spherical coordinates. As previously explained, SH defines a set of harmonic functions by solving the angular Laplace equation in spherical coordinates. In our method, we separately project the time series of the spherical angle vector $f(\theta, \phi)$ of local joints and the spatial pose onto a two-sphere.

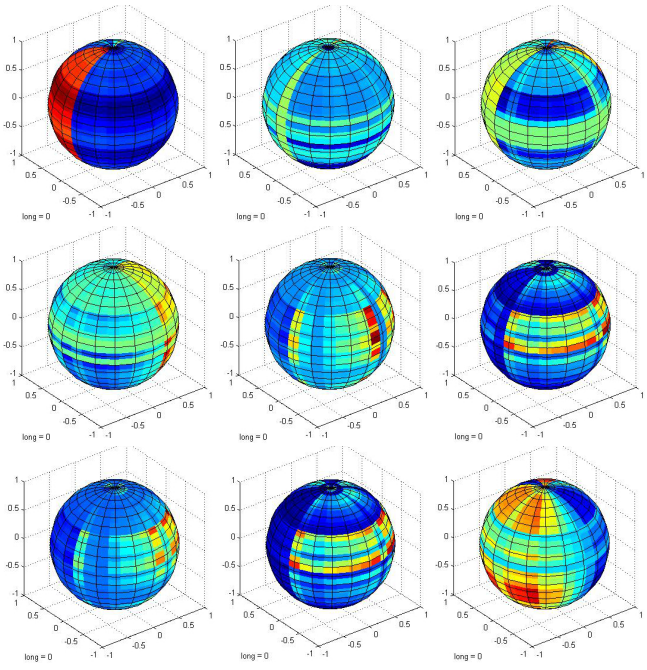

Fig. 3. Spherical harmonics basis are computed for local joints. The temporal motion of each local joint is mapped onto a unit sphere. The unit spheres in the plot represent individual joint of a person performs a tennis swing action. (Top panel, left to right):Elbow R/L, Wrist R/L, Knee Right. (Bottom panel, left to right) Knee left, Foot R/L , and Head Joints respectively

Let the entire skeleton body be represented by J joints (i.e., $J = (1, 2, , j)$, and the action be performed over T frames. Also, let $x_t(\rho, \theta, \phi) \in R^3$ denote the spherical coordinates of the skeleton joints in the human body at each frame. The spherical angle system of the entire action sequence can be constructed as a spatiotemporal system expressed below.

$$F_{s\in A}(\theta,\phi) = Pose\downarrow \begin{matrix} J_1 \\ J_2 \\ \vdots \\ J_k \end{matrix} \begin{bmatrix} (\theta,\phi)_{1,1} & (\theta,\phi)_{1,2} & \dots & (\theta,\phi)_{1,T} \\ \theta,\phi)_{2,1} & (\theta,\phi)_{2,2} & \dots & (\theta,\phi)_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ (\theta,\phi)_{J,1} & (\theta,\phi)_{J,2} & \dots & (\theta,\phi)_{J,T} \end{bmatrix}. \tag{9}$$

where s is the specific action, T is the total number of frames in the action sequence, and J is the number of joints in skeleton sequence.
The spatiotemporal system in 9 combine the spherical angles for temporal displacement(row) and spatial distribution (column) of local join, and pose in the action sequence, respectively.

However, using the system in equation (9 and MSHs concept, the static pose description can be calculated by projecting each column in equation 9 onto the basis function of MSHs (Silverman, 1972).

$$f(\omega_q) = \sum_{l=0}^{\inf} \sum_{m=-l}^{l} f_n^m,\ Y_n^m\ (\omega_q). \tag{10}$$

where $f(\omega_p) \in R$ is the real valued spherical angle vector. Recall that $\omega_q, q \in \{0, ..., j-1\}$ is the vector pointing at the angle of colatitude $\theta_q \in [0, \pi]$ measured down from the upper pole, and the angle of longitude $\phi \in [0, 2\pi]$ is the argument of the S^2 in spherical coordinates. The expansion coefficients are calculated using

$$f_n^m = 4\pi/n \sum_{\alpha=0}^{n-1} f(\omega_q)\ Y_n^m(\omega_q). \tag{11}$$

where $Y_n^m(\omega_q)$ is the modified real SH basis function defined as

$$Y_n^m(\omega_q) = \sqrt{2} Q_n^m\ [2\cos^2\phi - 1] Z_n^m(\cos\theta),\ m = 2. \tag{12}$$

In Eq. (12), Q is the scaling factor, and Z is the associated Legendre polynomials given as.

$$Z_n^m(x) = \frac{-1^m}{(2^n n!)}(1+x^2)^{\frac{m}{2}} \frac{(d^{n+m})}{(dx^{n+m})}(x^2-1)^n. \tag{13}$$

The estimated MSHs for the body pose at time t (column of Eq. 9) represent the spatial features of static pose.

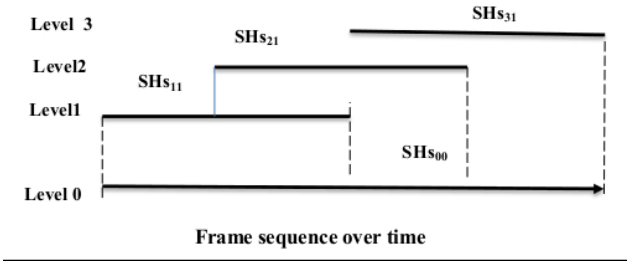

Fig. 4. A 3-level representation of Temporal construction of the SHs, $SHs_l j$ is the j^{th} Spherical harmonics in the l^{th} level of the hierarchy

The collection of the estimated MSHs for all poses of a specific action defines the spatial distribution feature vector of $\mathbf{H^s} = [\mathbf{P_1}, \mathbf{P_2}, ..., \mathbf{P_T}]$. Similar to defining static pose, the MSHs of the local joint displacement are calculated by projecting each row of Eq. 9 onto the basis function of MSHs. In this case, the individual MSH of each local joint displacement is calculated over the entire row. To compute the MSH feature vector of the local joints for a given action segment, we collect the individual motion vectors $\mathbf{H^m} = [\mathbf{M_1}, , \mathbf{M_J}]$.
AS an illustration example, the MSHs of a subject performing a tennis forehand action are depicted in fig. 3.2. In this figure, each panel demonstrates the harmonic distribution of the local joint motion displacement.

3.5. Temporal Construction of MSHs in a Hierarchical Model

In 3D skeleton-based action recognition, a compact skeleton-based descriptor should encode the static pose information and the temporal evolution or joint motion at a given time segment. The static pose and joint displacement features of a given skeleton body sequence contain discriminative data about the human action over a time segment.
In the previous section, the MSHs capture the spatial dependency of the holistic joints (i.e., pose in frame) and the motion of the local joint properties over the time sequence.
To efficiently encode the temporal variation of the local joints over time, each SH of these joints is constructed in a hierarchical manner. The idea of hierarchical construction is inspired by the spatial pyramid matching introduced by (Lazebnik et al., 2006) to achieve matching in 2D images. Relying on determining the MSHs calculated in the previous section, we construct the MSHs of the local joints in a multi-level approach. Each MSH covers a specific time window of the action sequence. The MSHs are computed over the entire video sequence from the top level and over the smaller windows at the lower levels. Window overlapping is used to increase the ability of the proposed representation to differentiate multiple actions by sliding from one window to the half of the next one, as depicted in Fig 4.
Regardless of whether the multiple levels of SHs are used, differentiating the local temporal sequences of various action categories is a difficult task because of numerous issues, including the frame rate variations and the temporal independence in each sublevel. To address these issues, DTW (Muller) is used to compute for a distance between the multiple levels of MSHs for each action category. DTW is a time series alignment algorithm. It compute a path between two sequences by warping the time axis iteratively until an optimal match between the two sequences is found. Similarly, DTW is used to identify the nominal distances between the MSHs of consecutive levels for each local joint. The distance vector for each local joint displacement is then formed. The temporal model of the skeleton joints is encoded for each action category as a concatenation of the distance vector $\mathbf{D^t} = [\mathbf{T_1}, , \mathbf{T_J}]$. Through the computation of the pose and motion feature vectors of the skeleton joints, an action sequence is represented by a combination of these vectors to form a skeleton representation feature vector.

$$\mathbf{S} = \delta \mathbf{H^s} + \kappa \mathbf{D^t}. \tag{14}$$

where δ and κ are the weighting parameters.
The static pose and temporal dynamic of the harmonics contain information about the spatiotemporal function over a time sequence of an action. Therefore, this type of harmonic information can be considered a compact representation of the body skeleton joint and can be used to reliably classify actions.

4. Action Classification

ELM is a multi-class classifier recently introduced for pattern recognition. The proposed action recognition system incorporates this classifier, which is a version of the feedforward neural network (Huang et al., 2006). Compared with other classifiers, ELM provides significant performances, such as fast learning time and recognition accuracy. In (Minhas et al., 2010), ELM was adopted for human activity recognition from video data. In recent years, this learning algorithm has been applied to solve skeleton-based human action recognition problems (Chen and Koskela) and many other computer vision problems. In this section, we present a brief review of the theory underlying this type of machine learning. For more details about the classical materials of ELM, see (Huang et al., 2006).

4.1. Extreme Learning Machine ELM

ELM has been extensively employed for learning single-hidden layer feedforward neural networks (Huang et al., 2006). The hidden nodes in ELM are randomly initialized and do not have to be iteratively tuned. In fact, these nodes remain fixed after initialization. As such, only the input weight parameters must be learned.
When the training sample A is given by (x_j, y_j), $j = [1, , q]$, in which $x_j \in R^N$ and $y_j \in R^M$, the output function of ELM model with L hidden neurons can be expressed as follows:

$$f_l(x) = \sum_{i=1}^{L} g_i \omega_i(x) = \mathbf{\Omega(x)}\, \mathbf{G}. \tag{15}$$

where $\mathbf{G} = [\mathbf{g_1}, \ldots, \mathbf{g_L}]$ is the output weight vector relating the L hidden nodes to the $m > 1$ output nodes, and $\mathbf{\Omega}(x) = [\omega_1(x), ..\omega_L(x)]$ is a nonlinear activation function (Huang et al.,

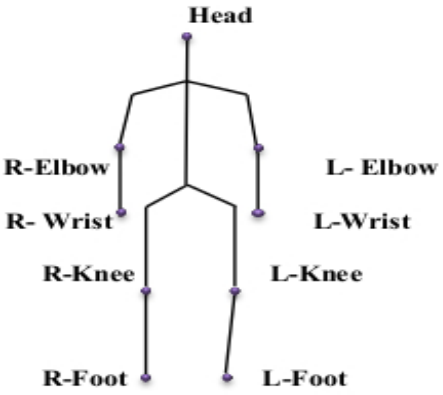

Fig. 5. Marked skeleton joints as captured by the Kinect sensor

2006). The system $\mathbf{\Omega}_i(x)$ can be written in an explicit form presented as follows:

$$\mathbf{\Omega}_i(x) = \beta(\tau_i.x + \epsilon_i), \tau_i \in R^d, \epsilon_i \in R. \tag{16}$$

where $\beta(.)$ is an activation function with hidden layer parameters (τ, ϵ). In the second stage of ELM learning, the error minimization between training data and output weight Ω is solved by using the least square norm depicted below.

$$min\|\mathbf{\Omega G} - \mathbf{H}\|^2, \mathbf{G} \in \mathbf{R}^{N*M}. \tag{17}$$

where$\mathbf{\Omega}$ defines the system of the layer of hidden neurons given as

$$\mathbf{\Omega} = \begin{bmatrix} \beta(\tau_1.x_1 + \epsilon_1) & \dots & \beta(\tau_L.x_1 + \epsilon_L) \\ \vdots & \ddots & \vdots \\ \beta(\tau_1.x_N + \epsilon_1) & \dots & \beta(\tau_L.x_N + \epsilon_L) \end{bmatrix}. \tag{18}$$

and $\mathbf{H}$ is the training data matrix denoted as

$$\mathbf{H} = \begin{bmatrix} \mathbf{h_1^T} \\ \vdots \\ \mathbf{h_N^T} \end{bmatrix}. \tag{19}$$

The optimal solution for minimizing the training error in (17) practically assumes that the number of hidden neurons L is less than that of the training set (i.e., $L < Q$). Therefore, in using the MoorePenrose generalized inverse of matrix $\mathbf{\Omega}$, the optimal solution for (17) is (Huang et al., 2012).

$$\mathbf{G}^* = \mathbf{\Omega}^*\mathbf{H}. \tag{20}$$

Where $\mathbf{\Omega}^*$ is the inverse of $\mathbf{\Omega}$.

4.2. Alternative Body Skeleton Features

Alternative skeleton representations are adopted as an anther abstraction of the skeleton features which are used for further performance evaluation of our method. These skeleton representations are as follows:

Joint Location (**JL**): simply concatenates all joint coordinates in one vector.

pairwise joints differences (**PJDs**): concatenation $y_f = \{p_i - p_j|i, j = (1, 2, , K), ij\}$ of all frames.

Magnitude of the Position Velocity (**MPV**): the velocity between the same joints of enter frame defined as $Y_{t1,t2} = \|p_{i,t1} - p_{i,t2}\|$.

These skeleton representations are fed directly into the classifier to directly compare the proposed MSHs method with the alternative representation schemes (i.e. JL, PJDs, and MPV).

5. Experimental Setup

To evaluate the effectiveness of the proposed method, we perform action recognition on the proposed feature representation and recently published datasets (i.e., MSR-Action 3D (Li et al., 2010),G3D (Bloom et al., 2012), Florence 3D Action (Seidenari et al., 2013), and UTKinect-Action (Xia et al., 2012)). These datasets are used as benchmarks in the experiment. The action complexity of these datasets varies from simple to complex sequences. In addition to depth data, skeleton data are also provided by these datasets using a Kinect sensor `http://msdn.microsoft.com/en-us/library/hh855352.aspx` as required.

In all experiments, an ELM classifier is used with the proposed representation. For each dataset, the state-of-the-art skeleton-based methods are extensively compared with the proposed approach. The number of hidden neurons of ELM is experimentally tuned for each dataset. To simplify the computation in each experiment, we set the waiting parameters as $\delta = 0.5$ and $\kappa = 0.5$. The frequency bands of MSHs are equal to 2. We consider the cross subject protocol for the test setting in all datasets. In particular, half of the subjects are used for training, and the other half for testing. We also divide the MSR-Action dataset into further subsets, **AS1, AS2** and **AS3** , similar to (Li et al., 2010). Each subset consists of eight actions, and action recognition is separately performed on each subset. In all experiments, we use nine joints from the body skeletal as the initial input joints to our proposed method, as shown in Fig 5. The features from these joints are initially used for the skeleton feature representation. All the results in this paper were reported in term of an average accuracy.

6. Results

6.1. Comparison with Various Skeleton Features

The performance of various skeleton representations is evaluated on all datasets, and the efficiency of the proposed MSHs

Table 1. Recognition rates (in %) for various skeletal representations on MSR-Action3D dataset

subset	JL	PDJ	MPV	Proposed method
AS1	72.2	76.22	80.23	**89.76**
AS2	69.83	80.47	79.15	**91.7**
AS3	82.7	71.4	84.06	**92.5**
Average	74.91	72.36	81.14	**90.98**

method is compared with the various skeleton representations. Table 1 reports the accuracy of the proposed approach with the corresponding results of different representation methods based on the MSR-Action dataset. Our findings presented in this table are achieved using three levels of MSHs, while the window overlap in the second and third levels is preserved. Compared with other skeleton representations, the proposed method provides satisfactory results. In particular, the MSHs method improves the average accuracies over JL, PDJs, and MPV by 16.07 %, 18.62 %, and 9.84 % , respectively. These observations clearly indicate the superiority of the proposed MSHs representation over existing skeleton representations. The experiments on MSR dataset indicate that the MSHs-based method have better performance on subsets AS1, AS2, AS3 in modeling complex actions and differentiating similar actions.

Table 2 summarizes the recognition accuracies of various skeleton representations on the UTKinect-Action, Florence 3D Action, and G3D datasets. The results reveal that MSHs method significantly outperforms the other skeleton representations on these datasets. In using UTKinect dataset, the average accuracy of the proposed representation is 10.5. % better than that of JL, 9.92% better than that of PJDs, and 5.42 % better than that of MPV. In the case of the Florence dataset, the average accuracy of the proposed representation is 9.54 % better than the average accuracy of, 15.8 % better than the average accuracy of PJDs, and 2.43 % better than the average accuracy of MPV. respectively. In the case of the G3D dataset, the average accuracy of the MSHs-based skeleton representation is 13.83 % better than the average accuracy of JL, 12.47 % better than the the average accuracy of PJDs, and 10.79 % better than the average accuracy of MPV. It is interesting to note that the results from MSHs clearly demonstrate the superiority of the MSHs-based method over various skeleton features on Florence, and G3D datasets.

6.2. Comparison with the state-of-the-art

The same datasets are used to compare the performance of the proposed method with those of existing state-of-the-art methods. For each data set, the hidden neurons are reported separately. In all experiments, the results correspond to using three levels of hierarchical SHs, while preserving the overlap in the last two levels.

Several recognition results on the MSR-Action 3D dataset are already available in the literature. Table 3 presents the recognition rate of the proposed approach along with those of the corresponding current methods. As indicated in this table, the proposed approach obtains the best results compared with those of most existing methods. In particular, our method provides good

Table 2. Recognition rates (in %)for various skeletal representations on UTKinect Action, Florence3D Action, and G3D Action datasets

Dataset	JL	PDJ	MPV	Proposed
UTKinect	82.5	83.08	87.58	**93.0**
Florence3D	76.59	70.33	83.7	**86.13**
G3D	79	80.36	82.04	**92.89**

Table 3. Comparison of recognition rates (in %) with the state-of-the-art results on MSR-Action3D dataset

Approaches	Accuracy
HO3DJ Xia et al. (2012)	78.97
EigenJoints Yang and Tian (2012)	82.30
Joint angles similaritiesOhn Bar and Trivedi (2013)	83.53
Fusion spatiotemporal Zhu et al. (2013)	90.90
Covariance Descriptor Hussein et al. (2013)	90.53
Skeletal Quads Evangelidis et al. (2014)	89.86
Lie Group Raviteja et al.	92.46
SHs Al Alwani and Chahir (2015)	90.94
proposed approach	**90.98**

results in line with those of some existing methods but outperforms the others. Moreover, The proposed approach generates the best results over the stat-of-the-arts reported in[Xia et al. (2012); Yang and Tian (2012); Ohn Bar and Trivedi (2013); Evangelidis et al. (2014)]. We can see that our result is in line with recent stat-of-the-arts such as [Zhu et al. (2013); Hussein et al. (2013); Al Alwani and Chahir (2015)], but is inferior to Raviteja et al.. This is probably because these methods either, incorporates a spatiotemporal skeleton joints structures with huge numbers of trees in the classification stage as in Zhu et al. (2013), or incorporates geometry relation between local joints in order to models the rotation and translation invariants of the features as lie groups as in Raviteja et al.. Some skeleton based methods like Yang and Tian (2012) use skeleton features based on pairwise differences between joints. However, results obtained on MSR Action 3D dataset show that integrating the evolution of the whole skeleton during the sequence is more discriminative than taking into consideration the joints separately. In addition, the methods proposed in Yang and Tian (2012); Xia et al. (2012); Ohn Bar and Trivedi (2013) have the lack of information about temporal nature of the action,making the recognition less effective compared to our method. In this case, 780 hidden neurons are observed in ELM.

For further evaluation, the proposed approach is applied to the skeleton sequences from UTKinect-Action, Florence, and G3D Action datasets. The performance of the proposed approach in this experiment is also compared with those of the corresponding methods. Table 4 compares our method with various state-of-the-art skeleton-based human action recognition approaches on the UTKinect dataset. The proposed approach gives comparable results. The average accuracy of the proposed representation is 5.10% better than that given in (Zhu et al., 2013) and 2.08% better than that in (Xia et al., 2012). The number of hidden neurons in this experiment is 640.

Table 5 reports the average recognition accuracies in the case of the Florence dataset. The results reveal that the accuracy of the proposed method is slightly higher than that citepd in (Seidenari et al., 2013). In particular, the performance of the proposed approach is superior over that of the state-of-the-art methods by 4.13%. Our results in this table correspond to 500 hidden neurons for ELM.

Table 4. Comparison of Recognition rates (in %)with the state-of-the-art results using UTKinect dataset

Fusion spatiotemporal Zhu et al. (2013)	87.90
HO3DJ Xia et al. (2012)	90.92
Space-time pose Rep. Devanne et al. (2013)	91.5
SHs Al Alwani and Chahir (2015)	91.65
Proposed approach	**93.0**

Table 5. Comparison of recognition rates (in %) with the state-of-the-art results, using Florence dataset

Multi-part bagSeidenari et al. (2013)	82.00
SHs Al Alwani and Chahir (2015)	87.50
Proposed approach	86.13

Table 6. Comparison of recognition rates (in %) with the state-of-the-art results, using G3D dataset

Bloom et al. (2012)2012	71.04
Alwani et al. (2014)	80.55
SHs Al Alwani and Chahir (2015)	92.30
Proposed approach	**92.89**

The performance of the proposed method is also assessed based on the G3D-Action dataset. Table 6 demonstrates the results, which indicate that our method evidently outperforms the existing skeletal joint-based state-of-the-art methods by achieving better accuracy by 0.59%. In this experiment, 700 hidden neurons exist in the ELM.

6.3. Benefit of modified SHs

Table 7 demonstrates that the addition of dynamic features expressed by the second-order term of the real SHs dramatically increases the recognition accuracy compared with the standard SHs (Al Alwani and Chahir, 2015). The efficiency of using MSHs becomes evident when we compare them with the standard SH descriptors. In Table 4, the recognition accuracies of MSHs are used and compared with those of the standard SHs given in (Al Alwani and Chahir, 2015). The explicit estimation of angular speed and directions in terms of the second-order function presents a significant performance. For example, in the MSR-Action 3D dataset, the use of the quadratic term in MSHs improves the recognition accuracy by a substantial .04% margin over the standard SHs. In the case of the UTKinect and G3D datasets, the MSHs add a significant improvement of 1.35% and 0.59% to their recognition accuracies respectively.
Contrarily, in the Florence dataset, the recognition rate is decreased from 87.5% for SHs to 86.13% for MSHs.
Our findings affirm that the angular speed component of the quadratic function in MSHs model is extremely important for action representation with curved displacement. Such a displacement cannot be fitted by the spatiotemporal features of the standard real SHs.

Table 7. Comparison of recognition rates (in %) with the SHs-based state-of-the-art results

Datasets	**SHs**	**MSHs**
MSR Action 3D	90.94	**90.98**
UTKnect	91.65	**93.00**
Florence	**87.50**	86,13
G3D	92.30	**92.89**

Table 8. Recognition rate (in %)for different classifiers

Dataset	**SVM**	**ELM**
MSR-action	86.36	**90.98**
UTKinect	89.94	**93.0**
Florence3D	79.06	**86.13**
G3D	91.5	**92.89**

6.4. Comparison with other Classifier

To further asses the performance of the proposed approach, we also compare the performance of ELM with that of a support vector machine (SVM) classifier and report the obtained recognition accuracies on the same datasets. We simply use a linear SVM (Chang and Lin, 2011) to compare the recognition algorithms.
The recognition accuracies corresponding to ELM and SVM are reported in Table 8. Our finding indicates that ELM performs better than SVM based on the MSR dataset, achieving a recognition accuracy of 90.98% (as opposed to the 86.36 % attained by SVM). Nonetheless, both classifiers exhibit distinctive results in the case of the UTKinect dataset. The performance of SVM based on the Florence dataset is slightly lower than that of ELM; the recognition rate of the latter is 6.07% higher than that of the former. For the G3D dataset, the result of ELM conforms to that of SVM.

7. Conclusion

In this study, we introduced a novel framework for action recognition based on an explicit model of 3D skeleton joints in a spatiotemporal domain. SH transform was used to explicitly model the angular speed and direction of joints. A novel MSH was also proposed based on the quadratic function of the real part of SHs. According to our framework, all body joints were registered into a body coordinate system to extract the spherical angles of joints expressed in 3D body coordinates. The spatiotemporal system of human action was then constructed and encoded by a set of MSHs of static poses and local joint displacement over time. The temporal evolution of the skeleton joints was characterized in a hierarchical manner. An appropriate skeleton representation of an action was formulated as a vector of combined poses and joint motion features. For the action recognition, ELM was used. The performance of the proposed method was evaluated based on recent 3D skeleton-based datasets. We compared the proposed method with the existing state-of-the-art methods either by adopting pure skeleton

data or by directly relying on depth data. The experimental results revealed that depending on the used datasets, the proposed method can obtain results similar to those of the extant methods or outperform them. The findings also indicated that MSHs are well suited for action representation with curved movement and angular direction changes.

In summary, our newly proposed method is in line with the recently presented methods for 3D skeleton-based pose representation. The angular direction estimated from skeleton data and its derived SHs are relevant for action recognition and can be successfully used to capture temporal changes in action and obtain a high recognition rate.

References

Aggarwal, J., Xia, L., 2014. Human activity recognition from 3d data: A review. Pattern Recognition Letters 48, 70–80. doi:10.1016/j.patrec.2014.011.

Al Alwani, A., Chahir, Y., 2015. 3-d skeleton joints-based action recognition using covariance descriptors on discrete spherical harmonics transform, in: International Conference on Image Processing (ICIP) 2015, IEEE, Québec, Canada.

Alnowami, M., Alnwaimi, B., Tahavori, F., Copland, M., Wells, K., 2012. A quantitative assessment of using the kinect for xbox360 for respiratory surface motion tracking, in: Proc. of SPIE, pp. 1–10.

Alwani, A.A., Chahir, Y., Goumidi, D.E., Molina, M., Jouen, F., 2014. 3d-posture recognition using joint angle representation, in: Information Processing and Management of Uncertainty in Knowledge-Based Systems - 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part II, pp. 106–115. doi:10.1007/978-3-319-08855-6_12.

Antnio, W.V., Thomas, L., William, S., Mario, F.M.C., 2012. Distance matrices as invariant features for classifying mocap data, in: ICPR 2012 (21st International Conference on Pattern Recognition), IEEE. pp. 2934–2937.

Barnachon, M., Bouakaz, S., Boufama, B., Guillou, E., 2013. A real-time system for motion retrieval and interpretation. Pattern Recognition Letters 34, 1789–1798. doi:10.1016/j.patrec.2012.12.020.

Bloom, V., Makris, D., Argyriou, V., 2012. G3d: A gaming action dataset and real time action recognition evaluation framework., in: CVPR Workshops, IEEE. pp. 7–12.

Bustos, B., Keim, D.A., Saupe, D., Schreck, T., Vrani, D.V., 2005. Feature-based similarity search in 3d object databases. ACM Computing Surveys 37, 345–387.

Chang, C.C., Lin, C.J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, X., Koskela, M., . Skeleton-based action recognition with extreme learning machines, in: (In press) Neurocomputing. doi:10.

Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Bimbo, A.D., 2013. Space-time pose representation for 3d human action recognition., Springer. pp. 456–464.

Evangelidis, G., Singh, G., Horaud, R., 2014. Skeletal quads: Human action recognition using joint quadruples, in: International Conference on Pattern Recognition, Stockholm, Sweden. pp. 4513 – 4518.

Freeden, W., Michael, S., 2009. Spherical Functions of Mathematical Geosciences: A Scalar, Vectorial, and Tensorial Setup. First ed., Springer-Verlag Berlin Heidelberg.

Huang, G.B., Zhou, H., Ding, X., Zhang, R., 2012. Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B 42, 513–529.

Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70, 489 – 501. doi:http://dx.doi.org/10.1016/j.neucom.2005.12.126.

Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M., 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press. pp. 2466–2472.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, p. 21692178.

Li, W., Zhang, Z., Liu, Z., 2010. Action recognition based on a bag of 3d points, in: CVPR Workshop, IEEE. pp. 9–14.

Lv, F., Nevatia, R., 2006. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, in: Proceedings of the 9th European Conference on Computer Vision - Volume Part IV, pp. 359–372.

Minhas, R., Baradarani, A., Seifzadeh, S., Jonathan, W.Q., 2010. Human action recognition using extreme learning machine based on visual vocabularies. Neurocomputing 73, 1906–1917.

Muller, M., . Information Retrieval for Music and Motion.

Obdrzalek, S., Kurillo, G., Ofli, F., Bajcsy, R., Seto, E., Jimison, H., Pavel, M., 2012. Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE. pp. 1188 –1193. doi:10.1109/EMBC.2012.6346149.

Ohn Bar, E., Trivedi, M., 2013. Joint angles similarities and hog2 for action recognition, in: CVPRW, pp. 465–470.

Raptis, M., Kirovski, D., Hoppe, H., 2011. Real-time classification of dance gestures from skeleton animation, in: Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 147–156.

Raptis, M., Wnuk, K., Soatto, S., 2008. Flexible dictionaries for action classification, in: The 1st International Workshop on Machine Learning for Vision-based Motion Analysis - MLVMA'08.

Raviteja, V., Felipe, A., Rama, C., . Human action recognition by representing 3d skeletons as points in a lie group, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE.

Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A., Pala, P., 2013. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 479–485.

Shimada, A., Taniguchi, R.i., 2008. Gesture recognition using sparse code of hierarchical som., in: ICPR, IEEE Computer Society. pp. 1–4.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., . Real-time human pose recognition in parts from a single depth image, in: CVPR, IEEE.

Silverman, R.A., 1972. Special Functions and their Applications. Dover Publications.

Vranic, D., 2003. An improvement of rotation invariant 3d-shape descriptor based on functions on concentric spheres, in: ICIP03, IEEE. pp. 757–760.

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras., in: CVPR, IEEE Computer Society. pp. 1290–1297.

Wang, J.Y., Lee, H.M., 2009. Recognition of human actions using motion capture data and support vector machine. WRI World Congress on Software Engineering, WCSE 1, 234–238. doi:10.1109/WCSE.2009.354.

Xia, L., Chen, C.C., Aggarwal, J.K., 2012. View invariant human action recognition using histograms of 3d joints., in: CVPR Workshops, IEEE. pp. 20–27.

Yang, X., Tian, Y., 2012. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor., in: CVPR Workshops, IEEE. pp. 14–19.

Yao, A., Gall, J., Fanelli, G., Gool, L.V., 2011. Does human action recognition benefit from pose estimation?, in: Proceedings of the British Machine Vision Conference, BMVA Press.

Ye, M., Zhang, Q., 0002, L.W., Zhu, J., Yang, R., Gall, J., 2013. A survey on human motion analysis from depth data., in: Grzegorzek, M., Theobalt, C., Koch, R., Kolb, A. (Eds.), Time-of-Flight and Depth Imaging, Springer. pp. 149–187.

Zhang, L., Samaras, D., 2006. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 351–363. doi:http://doi.ieeecomputersociety.org/10.1109/TPAMI.2006.53.

Zhao, X., Li, X., Pang, C., Wang, S., 2013. Human action recognition based on semi-supervised discriminant analysis with global constraint. Neurocomputing 105, 45–50. doi:10.1016/j.neucom.2012.04.038.

Zhu, Y., Chen, W., Guo, G., 2013. Fusing spatiotemporal features and joints for 3d action recognition, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 486–491.