3-D skeleton joints-based action recognition using covariance descriptors on discrete spherical harmonics transform

Adnan Al Alwani, Youssef Chahir

▶ **To cite this version:**

Adnan Al Alwani, Youssef Chahir. 3-D skeleton joints-based action recognition using covariance descriptors on discrete spherical harmonics transform. International Conference on Image Processing (ICIP 2015), IEEE, Sep 2015, Québec, Canada. hal-01168436

HAL Id: hal-01168436
https://hal.science/hal-01168436v1

Submitted on 25 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

3-D SKELETON JOINTS-BASED ACTION RECOGNITION USING COVARIANCE DESCRIPTORS ON DISCRETE SPHERICAL HARMONICS TRANSFORM

Adnan AL ALWANI, Youssef CHAHIR

GREYC CNRS (UMR 6072)
University of Caen Basse-Normandie
{adnan.alalwani, youssef.chahir}@unicaen.fr

ABSTRACT

In this paper, we explore a new method for skeleton-based human action recognition. First, the normalized angles of local joints are extracted and Spherical Harmonics Transform (SHT) can then be used to explicitly model the angular skeleton by projecting the spherical angles onto unit sphere basis. This enables that the skeleton representation can be decomposed into a basis functions. We adopt the spatiotemporal covariance matrix of the spherical harmonic to capture joints orientations over the human action sequence. Thus, the covariance coefficients of joints are used as a discriminative descriptor for the sequence. We validate the proposed method using Extreme Learning Machine (ELM) classifier and recent published 3D action datasets. Experimental results show that our method performs better than many classical methods.

Index Terms—
Spherical Harmonics Transform, Skeleton Joints, Covariance Descriptor, Extrem Learning Machine

1. INTRODUCTION

Human action can be captured by either 2D RGB camera or recent release of 3D sensor. In the former case the holistic or body parts of human are used for human action recognition. however Despite the finding of these approaches, the monocular RGB data is still restricted by the various factors like shading, body part occlusion and background noise. Moreover, 2D video sensors cannot fully register the 3D motion of the human using single camera. On the other hand, the 3D information captured by RGB-Depth (RGB-D) sensor are used recently for human action recognition. The RGB-D such as kinect sensor provides both the 2D image as well as the depth map.

According to the motivation of [5] and the articulated structure of human body, human actions can be abstracted by a set of 3-D joints locations of poses. In which, the detection of the human posture by means of skeleton joints is practically achieved through further processing of the depth data. Based on the study of [24] , it obvious that using joints locations alone may provides a good human motion representation for action recognition task. However, in the field of 3D-based action recognition the trend has been focused on using the depth and skeleton joints data to develop an efficient methods for specific task recognition [22], and computer vision applications, etc... To this end, features estimation from skeleton body joints is relatively faster due to low dimensionality of the features vector and encodes a better view invariance.

This paper attempts to address the sequences of skeletal-joints representation problem in an explicit model. In this model, a novel feature descriptor is used based on the Spherical Harmonic Transform (SHT) of temporally local joints and the covariance coefficients. The main objective of our approach is based on the calculation of the SHT of spherical angles of local joints to explicitly model the displacement of each individual joint. Unlike the works in [3, 10, 13] that consider the spatial-relation between individuals joints. While the present study is related to recent approaches in skeleton descriptor [3], it capitalizes on a new feature space, which was not considered in these earlier studies.

Let a spherical coordinates of skeleton joint Ji denoted by (θ, ϕ), the model of temporal evolution ofJi can be represented using spherical harmonic of θ and ϕ orientation respectively. Then, to handle frame length variations, for each action category, we introduce the covariance technique to compute the covariance coefficients of each SHs matrix. Collecting the computed covariance coefficients of all selected local joints forms the skeleton features representation for an action sequence. Finally, we perform the evaluation procedure using ELM classifier and multiple 3D action datasets.

2. RELATED WORKS

With advance imaging techniques, such as Microsoft Kinect, an action recognition is abstracted by a set of simple extracted and low dimensional features. Moreover, using skeleton joints locations captured by this sensor, the body pose estimation, as well as action recognition can be achieved efficiently. In this section, we summarize various methods that only use skeleton data, and more related to our approach. A

rich material of the human motion analysis from depth data, can be found in [1].

Human skeleton was represented in [3] using 3D skeleton joint locations and the temporal evolutions were modeled using a temporal hierarchy of covariance descriptors. In [6], 3D coordinates of joints were used and the action sequence was modeled with a generative discrete HMM. Action recognition was performed using multi-class Adaboost. The proposed work in [17] used the idea of pairwise relative locations of the joints in order to represent human skeleton. The temporal displacement were characterized using a coefficients of Fourier pyramid hierarchy. The authors proposed an actionlet-based approach in which learning kernel approach was used in order to effectively candidates the meaningful joint combinations. In [[20]], the authors adopted a representation based on eigenjoint descriptor calculated from each frame. The action recognition was performed using the Naive-Bayes nearest neighbor. In [12] the task of action recognition is achieved by random forests classifier. A view invariant representation of human skeleton was proposed in [18] 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. Then, a generative HMMs classifier classifies each visual code word. Besides that, the authors of [13] used the idea of the skeletal quad to encodes the relation of local joint in a quadruples form. The skeletal quads are generated by a Fisher kernel representation based on Gaussian mixture model. In [10] a human skeleton was represented as points in the Lie group. The proposed representation explicitly models the 3D geometric relationships between various body parts, using rotations and translations. Since the Lie group is a curved manifold, they map all the action curves from the Lie group to its Lie algebra and the temporal evolutions were modeled using dynamic time warping (DTW).

3. SHPERICAL HARMONICS

In this section, we briefly discuss the spherical harmonics transform, the classical introduction of SHs can be found in[14]. Spherical harmonics are the solution to a variety of problems that relay on an orthonormal basis s^2. SHs was used for solving PDEs in geophysics, quantum mechanics, as well as a host of computer vision and computer graphics related applications [16, 11].

Suppose that the SHs denoted by y_l^m, are the angular solution that satisfies Laplace's formula in spherical coordinates: radial distance $r \in \mathbb{R}^+$, azimuth angle $\theta \in [0, 2\pi]$ and elevation angle $\phi \in [0, \pi]$. Then, the standard SHs have the expression [14]:

$$Y_l^m(\theta, \phi) = K_l^m P_l^m \cos(\theta) \exp(jm\phi). \quad (1)$$

Where $P_l^m cos(\theta)$ are the associated Legendre polynomials of degree l and order$m$, defined by the differential equation as

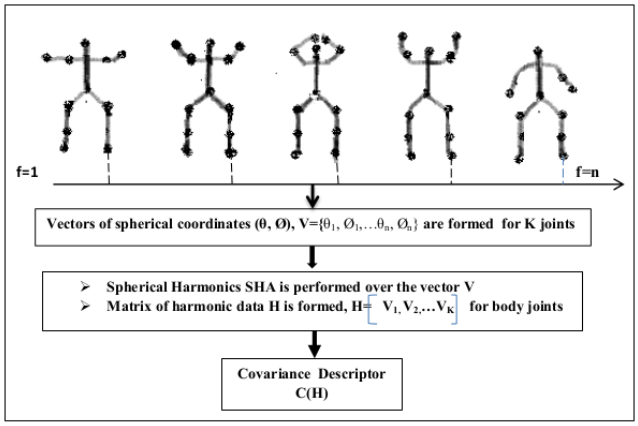

Fig. 1. Construction of the 3D joints descriptor

$$P_l^m = \frac{-1^m}{(2^l l,\ !)}(1+x^2)^{\frac{m}{2}} \frac{(d^{l+m})}{(dx^{l+m})}(x^2-1)^l. \quad (2)$$

The term K_l^m is normalization constant, equal to

$$K_l^m = \sqrt{\left(\frac{(l+1}{4m}\right) \frac{(l-|m|)!}{(l+|m|)!}}. \quad (3)$$

Since the spherical harmonics is analogue to Fourier series on the unit sphere, any function $f$may be defined by a set of a linear combination of the harmonics basis. For this , the function $x(\theta.\phi)$ is decomposed in terms of the spherical harmonics as :

$$x(\theta, \phi) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} C_l^m Y_l^m(\theta, \phi). \quad (4)$$

Where C_l^m is the expansion coefficients, and $Y_l^m(.)$ is the real coefficients of the spherical harmonic given by :

$$Y_l^m(\theta, \phi) = \sqrt{2K_l^m} \cos(m\phi) P_l^m(x), \quad \text{for } m > 0. \quad (5)$$

4. 3D POSE DESCRIPTOR

4.1. Overview

Regardless of the skeleton structure being used, temporal sequence discrimination into different action classes is a difficult task due to challenges like frame numbers variations in each action, and temporal joints dependency. To address these problems for each action class, we propose a highly discriminative 3D pose descriptor. Particularly, we introduce a novel skeleton-joints descriptor that is based on covariance between local joints. As shown in Fig. 1, the descriptor is designed by finding the covariance coefficients on the spherical harmonics

of local joints. We sample these coefficients over the time of the action sequence.

The idea of covariance descriptor was first adopted by [7] as a region descriptor of an image and texture-based classification [8]. The idea of spatio-temporal patch-based covariance descriptor is recently introduced as an action recognition framework [2]. In our work, we compute the spatio-temporal covariance coefficients between local joints along the time sequence.

4.2. Covariance-based descriptor for body skeleton joints

Suppose we have the entire skeleton structure is represented by Q joints, and the action is performed over T time sequence (frame). Let H denote harmonics data matrix of a set of spherical harmonics $\{h_1, ..h_n\}$. Because sets of related spherical harmonics of Q joints are considered for whole action, the 2-D SHs h_i of length $m = v \times u$ is expressed in column vector i.e. $= vect(h)$. Thus, the harmonic data H is an$M \times Q$ matrix, and defined as $\mathbf{H} = \{\mathbf{h_1}, ..., \mathbf{h_Q}\}$ where typically,$M > Q$ with fixed Q. Having obtained the harmonic data matrix$H$, the covariance elements over the sequence T is given by [3] :

$$C(H) = \frac{1}{T-1}\sum_{t=1}^{T}(H - \bar{H})(H - \bar{H}). \tag{6}$$

Where $\bar{H}$ is the sample mean of H.

In our case, we sample the lower part elements of the covariance matrix $C(.)$. Thus, the length of the descriptor is $Q(Q+1)/2$. Where Q is the number of skeleton joints used to represent the action sequence.

5. ACTION CLASSIFICATION

We employ Extreme Learning Machine ELM for the action classification. Recently, this learning algorithm has been applied to solve recent skeleton-based human action recognition problem [23]. ELM has been extensively devoted for learning single hidden layer feedforward neural networks SLFNs [21]providing fast learning time and accurate results.
We start the learning stage by assuming that there are M actions $A = A_1, ...A_M$ and the row vector $y = [y_1, ...y_M]$ indicates the action that the sequence belongs to. Note that, each action sequence c is represented by the features of its frames, calculated as described in section 4 i.e (c, y) form a set of training pairs for the classifier.

For the training samples $P\{x, y\}$ where $x_i \in R^n$ and $y_i \in R^m$, the output function of ELM model with N hidden neurons can be represented as [21] :

$$f_n(x) = \sum_{i=1}^{N}\omega_i\psi_i(x) = \mathbf{\Psi(x)\Omega}. \tag{7}$$

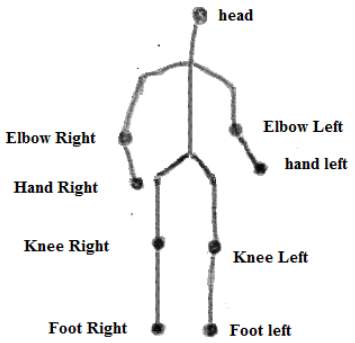

Fig. 2. Marked skeleton joints as captured by the Kinect sensor

Where $\Omega = [\omega_1, , \omega_N]$ is the output weight vector relate the N hidden nodes to the $m > 1$ output nodes, and $\Psi(x) = [\psi_1(x), ..\psi_N(x)]$ is a nonlinear activation function [21]. In particular, the system $\psi_i(x)$ can be written in explicit from as :

$$\psi_i(x) = \beta(\tau_i.x + \epsilon_i), \tau_i \in \mathbb{R}^d, \epsilon_i \in \mathbb{R}. \tag{8}$$

Where $\beta(.)$ is a mapping function, with hidden layer parameters (τ, ϵ). In the second stage of ELM learning, the errors between training data and the output weight Ω, is solved by minimizing the solution of the following term

$$min\|\mathbf{\Psi\Omega} - \mathbf{T}\|^2, \Omega \in \mathbb{R}^{N*M}. \tag{9}$$

WhereΨ define the system of the hidden neurons layer given as

$$\mathbf{\Psi} = \begin{bmatrix} \psi(\mathbf{x_1}) \\ \vdots \\ \psi(\mathbf{x_N}) \end{bmatrix}. \tag{10}$$

And $\mathbf{T}$ is the training data matrix denoted as

$$\mathbf{T} = \begin{bmatrix} \mathbf{t_1^T} \\ \vdots \\ \mathbf{t_N^T} \end{bmatrix}. \tag{11}$$

Using the Moore-Penrose generalized inverse of matrix $\mathbf{\Psi}$, the optimal solution to (9) can be found as [21].

$$\mathbf{\Omega}^* = \mathbf{\Psi}^*\mathbf{T}. \tag{12}$$

Where $\mathbf{\Psi}^*$ denotes the inverse of $\mathbf{\Psi}$.

6. EXPERIMENTS

In this section, quantitative results of action recognition are reported and compared. Four public RGB-D datasets, acquired using a Kinect sensor, were used as benchmarks in the experiment. These datasets are: MSR-Action3D

dataset [4], UTKinect-Action dataset [18], Florence3D-Action dataset[19] and gaming G3D dataset [9]. In all experiments, we used a ELM classifier with the covariance descriptor. Before computing the descriptor, we first need to project the 3D joints data into a common coordinate system to make the joints' coordinates. To achieve this, we select the hip-center as the origin point, and use its coordinates as the common basis. Then, we project and transform all the other skeleton joints on the new center.

Evaluation Settings and Parameters : For MSR-Action3D dataset, the protocol of cross subject test setting was used similar to [4]. We further divided the dataset into subsets AS1, AS2 and AS3 each consisting of 8 sub-actions. The recognition task was performed on each subset separately and we averaged the results. For the remaining data sets, we divide each dataset into half of the subjects for training and the rest are used for the testing task. We selected nine joints from the body skeletal as shown in Figure 2. These joints were used as an initial features input for descriptor. The number of hidden neurons were selected by experiment to perform high accuracies and our results are compared with state- of-the-arts methods that rely only on the skeleton joints description.

7. RESULTS

Previous recognition results have already been reported in the literature using the MSRAction3D dataset. Table 1 shows the recognition rate per action subset along with the corresponding results of methods that rely on skeleton joints. As we can see, our method gives a good results. More specifically, our method outperforms some of the state-of-the-art methods on this dataset.

Similar to [12], we experimented with our approach on a UTKinect-Action and Florence3D-Action datasets, and we use the same setup in [9] on a G3D Action dataset.

Table 1. Comparison of Recognition rates with the state-of-the-art results on MSR action dataset

Histograms of 3D joints [18]	78.97
EigenJoints [20]	82.30
Joint angle similarities [22]	83.53
Covariance descriptors [3]	90.53
Random forests [12]	90.90
Joints as special Lie algebra [10]	92.46
Proposed approach	**90.94**

Table 2. Comparison of Recognition rates with the state-of-the-art results using UTKinect dataset

Random forests [12]	87.90
Histograms of 3D joints [18]	90.92
Proposed approach	**91.65**

Table 3. Comparison of Recognition rates with the state-of-the-art results, using Florence dataset

Multi-Part Bag-of-Poses [19]	82.00
Joints as special Lie algebra [10]	90.88
Proposed approach	87.50

Table 4. Comparison of Recognition rates with the state-of-the-art results, using G3D dataset

Hybrid joints feature + adaboost [9]	71.04
Proposed approach	**92.30**

Table 2 summarizes the recognition accuracies of our method compared with current skeleton-based method using UTKinect dataset. In this case the proposed approach gives the best results on these datasets. For example, the average accuracy of our method outperforms the average accuracy of [18] and [12] by 0.73% and 3.75%, respectively.

We further evaluate our method using Florence dataset, the recognition rates compared with various methods were reported in table 3, The proposed method gives the best over the results of [19] by 5.5%.

The last experiment was carried out on G3D-Action dataset. The average accuracy of our representation reported in table 4 is 21.26%. This result is better than the average accuracy of [9]. These results clearly demonstrate the performance of our proposed method over a number of existing skeletal joints-base approaches.

8. CONCLUSION

The problem of skeleton body representation was explicitly modeled in this paper. We have presented an efficient approach for skeleton-based human action recognition. We adopted the spherical harmonics and covariance technique. We used the spatio-temporal spherical harmonics that characterize the spherical angles of local joints over the entire action sequence. We exploited the idea of covariance components in order to capture the dynamic of the action and provide a relevant descriptor with the a fixed length.

The experimental results tested on a various datasets prove the effectiveness of the proposed method. Results demonstrate that our method can be successfully used for capturing temporal changes in action and achieve a higher recognition rate. In future studies, we will enhance our method for classifying and recognizing different other behaviors.

9. REFERENCES

[1] J. Aggarwal and L. Xia, Human activity recognition from 3d data: A review," *pattern Recognition lettersl*, 2014.

[2] A. Sanin, C. Sanderson,M. Harandi, and B.C. Lovell, Spatio-temporal covariance descriptors for action and gesture recognition," *WACV*,2013.

[3] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban, "Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations," *IJCAI*,2013.

[4] W. Li, Z. Zhang, and Z. Liu", Action Recognition Based on a Bag of 3D Points," *In CVPRW*, 2010.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, Real-Time Human Pose Recognition in Parts from a Single Depth Image,"*(*In CVPR,2011.

[6] F. Lv and R. Nevatia, Recognition and Segmentation of 3D Human Action Using HMM and Multi-class Adaboost," *In ECCV*,2006.

[7] O. Tuzel, Fatih Porikli, and Peter Meer, Region covariance: a fast descriptor for detection and classification," *In ECCV*, pp. 589-600, 2006.

[8] O. Tuzel, F. Porikli, and P. Meer, Pedestrian detection via classification on Riemannian manifolds," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.30, pp.1713-1727, 2008.

[9] V. Bloom, D. Makris, V. Argyriou, "G3D: a Gaming action dataset and real time action recognition evaluation framework," *In CVCGW*, 2012.

[10] R. Vemulapalli, F. Arrate and R Chellappa, Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," *In ICCV*, 2014,

[11] B. Bustos, D. A. Keim, D. Saupe, T. Schrek, D. V. Vranic, Feature-based similarity search in 3-D object databases," *ACM Comp. Surveys*, Vol.37, pp. 345-387, 2005.

[12] Y. Zhu, W. Chen, and G. Guo, ""Fusing Spatiotemporal Features and Joints for 3D Action Recognition," *In CVPRW*, 2013.

[13] G. Evangelidis, G Singh and R. Horaud, "Skeletal Quads: Human Action Recognition Using Joint Quadruples," *In ICPR*, 2014.

[14] W. Freeden and M. Schreiner, "Spherical Functions of Mathematical Geosciences ; A Scalar, Vectorial, and Tonsorial Setup," *Springer Publisher*, 2009.

[15] M. Raptis, D. Kirovski, and H. Hoppes, Real-time classification of dance gestures from skeleton animation," *InSymp. on Comp. Anim*, 2011.

[16] L. Zhang and D. Samara, Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *IEEE Trans. PAMI*, Vol. 28, pp. 351-363, 2006.

[17] J. Wang, Z. Liu, Y. Wu, and J. Yuan, Mining Actionlet Ensemble for Action Recognition with Depth Cameras," *In CVPR*, 2012.

[18] L. Xia, C. C. Chen, and J. K. Aggarwal, View Invariant Human Action Recognition Using Histograms of 3D Joints," *In CVPRW*, 2012.

[19] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala, Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses," *In CVPRW*, 2013.

[20] X. Yang and Y. Tian, EigenJoints-based Action Recognition Using Nave-Bayes-Nearest-Neighbor," *In CVPRW*, 2012.

[21] G. Huang, H. Zhou, X. Ding, and R. Zhang, Extreme learning machine for regression and multiclass classification," *IEEE Trans. on, Systems, Man, and Cybernetics, Part B:Cybernetics*, Vol. 42, pp. 512-529, 2012.

[22] E. Ohn-bar and M. M. Trivedi, "Joint Angles Similarities and HOG2 for Action Recognition," *In CVPRW*, 2013.

[23] X. Chen, and M. Koskela,"Skeleton-Based Action Recognition with Extreme Learning Machines," *International Conference on Extreme Learning Machines, (ELM2013)*, Beijing, October, 2013.

[24] A. Yao, J. Gall, G. Fanelli, and L. Van Gool."Does human action recognitio benefit from pose estimation?." *In Proceedings of the British Machine Vision Conference*, pages 67.167.11, 2011.