Bayesian non-parametric inference for manifold based MoCap representation

Fabrizio Natola, Valsamis Ntouskos, Marta Sanzari, Fiora Pirri
ALCOR LAB, Dipartimento di Ingegneria Informatica Automatica e Gestionale
{natola,ntouskos,sanzari,pirri}@dis.uniroma1.it

Abstract

We propose a novel approach to human action recognition, with motion capture data (MoCap), based on grouping sub-body parts. By representing configurations of actions as manifolds, joint positions are mapped on a subspace via principal geodesic analysis. The reduced space is still highly informative and allows for classification based on a non-parametric Bayesian approach, generating behaviors for each sub-body part. Having partitioned the set of joints, poses relative to a sub-body part are exchangeable, given a specified prior and can elicit, in principle, infinite behaviors. The generation of these behaviors is specified by a Dirichlet process mixture. We show with several experiments that the recognition gives very promising results, outperforming methods requiring temporal alignment.

1. Introduction

Human action recognition is still a challenging and stimulating problem especially when considering motion capture data (MoCap), which are relevant in several applications including robotics, sports, rehabilitation and entertainment. A considerable amount of work has been proposed so far to solve problems arising in action recognition, such as view-point change, occlusions, likewise variations in behaviors amid different subjects performing the same action. However there is a significant difference between MoCap and 2D/2.5D action representations, and it could be argued without fear that the two recognition problems are drastically different, as they address different feature spaces and representations and, consequently, different recognition methods. MoCap sequences represent actions by 3D points, and joints of the human skeleton with appropriate kinematics. These data can, for example, be acquired by means of an RGB-D sensor, such as the Kinect, by infrared marker tracking systems, such as the Vicon System, or via back-projection techniques using multiple cameras. With this kind of data, occlusions so far have not been considered a major issue, such as with 2D/2.5 D data, however variations amid behaviors are still a major problem to be handled. Among the most relevant approaches we recall [14, 19, 17, 23, 30], all using noise and occlusion free datasets. In [14] actions are represented as structured-time series, with each frame lying on a high-dimensional ambient space, from which a spatio-temporal manifold is obtained by a dimensionality reduction approach, based on dynamic manifold warping, accounting only for joints translation. In [29], instead, both joints rotations and translations are considered, so as to construct a novel class of features in $SE(3)\times\cdots\times SE(3)$, obtaining a full feature space mapped on the Lie algebra. In [17] actions are represented via joint covariance descriptors, so as to work with symmetric positive definite matrices, which lie on Riemannian manifolds. In most of the approaches the representation of the joints space is a major issue and the need for a viable compromise between space reduction and completeness seems evident. In this sense we propose a novel representation for MoCap data, by introducing a new skeleton model, which has the advantage of considering the ambient space of the joints and mapping it into a reduced space via Principal Geodesic Analysis. The advantage of the proposed representation is that it keeps the most from the joints information and, at the same time, it provides the most suitable transformation to approach the recognition problem with a non-parametric Bayesian model.

Indeed, the representation model is crucial, both for eliciting features and for the recognition method used. For example, [14, 29, 19] consider a time-based ordering for which a temporal alignment is needed. In particular, [19] decompose the 3D joints into subspaces representing either the motion of a single body part, or of the combination of multiple ones. In our approach, instead, for each joint of the skeleton, and for each configuration in the action space, we keep the global transformation of the joint reference frame with respect to the world inertial frame. These transformation matrices are elements of a Riemannian manifold, and joints of the human skeleton have ranges of variation, which can be gathered into groups. In particular, we consider 6 sub-body groups, corresponding to the head, left and right legs, torso, left and right arms, respectively. Each of these defined groups represents a set of possible motions of the

associated sub-body part, and it is such that the elements in the set are order independent and exchangeable, making unnecessary the temporal alignment, as for example proposed in [14, 15, 29]. We provide a representation for these groups via the principal directions of each of them, in the configuration space. The obtained feature space proves to be good for classification, based on clustering. The basic idea is that every type of action generates specific set of behaviors for each sub-body part. To capture similarities amid behaviors we approach the classification problem with the Dirichlet process mixture model. Other approaches considering behaviors classification are [23, 30, 31]. In [23], the most informative joints are extracted by considering the fastest joints or the joints that mostly vary in angles. Similarly, [30] construct an actionlet ensemble, which is a collection of the most discriminative primitive actions, which in turn are the representative features of subsets of joints of an action sequence. These actionlets are learned within the SVM framework. [31] introduce eigenjoints as novel features so as to represent an action as the set of static pose, offsets and joints motion. Many approaches use datasets like [13, 7, 25, 24], which consider only 3D joints locations. Our approach, requiring full 3D poses, can be applied to these datasets too. In fact, following [29] the root joint (see Fig.1) can be simply considered translated with respect to the world origin, without rotations, and each other joint rotation matrix can be evaluated as the minimum rotation required to carry the world's x-axis onto the joint's bone.

The advantage of our approach is that behaviors are generated by Dirichlet process mixtures, exhibiting a great flexibility, and performing well both with queries formed by a single frame and with queries formed by a set of frames which do not need to be ordered, in so showing to be robust with respect to frame occlusions, actions interruptions, and looping repetitions. Indeed, the great benefit of the proposed method, called PGA-DPM, is that it provides a simple representation for basic actions, which is very suitable for learning. It can be used to generalize the recognition problem when time and subsequence relations are effectively needed to define complex actions, by combining different basic actions.

The paper is organized in the following manner. In Section 2, we focus on some preliminary definitions and methods that will be used to define the feature space. How groups of joints are obtained by collecting these features into groups, according to the limbs of the human skeleton, is explained in Section 3. In Section 4, we introduce the classification model based on Dirichlet process mixtures generating a representation of an action, which can possibly exploit some empirical knowledge of the action itself. In Section 5 results are presented, and a comparison with a state of the art method (the Dynamic Manifold Warping, [15, 14]) is proposed. Finally, in Section 6, we address some future developments together with some conclusive discussion.

2. Preliminaries

In this preliminary part we provide some basic notions that are used for the feature space representation, for further details on the basic concepts we refer the reader to [26, 11]. In the following, vectors are denoted by boldface symbols and matrices by upper case letters. We start considering the set of transformations T in $SE(n)$, $n = 3$:

$$T = \begin{bmatrix} R & \mathbf{d} \\ \mathbf{0}^{1\times 3} & 1 \end{bmatrix} \tag{1}$$

Here $R \in SO(3)$ is the rotation matrix, and $\mathbf{d} \in \mathbb{R}^3$ is the translation vector. $T \in SE(3)$ has 6 DOF and is used to describe the pose of the moving body with respect to the world inertial frame. $SO(3)$ and $SE(3)$ are Lie groups and their identity elements are the 3×3 and 4×4 identity matrices, respectively. The *tangent space* of a Lie Group at its identity element defines its *Lie algebra*. The Lie algebra $so(3)$ of $SO(3)$ is formed by skew-symmetric matrices of the form:

$$so(3) = \{\Omega \,|\Omega \in \mathbb{R}^{3\times 3}, \Omega = -\Omega^\top\} \tag{2}$$

Ω can be uniquely identified with a vector $\boldsymbol{w} \in \mathbb{R}^3$. The Lie algebra $se(3)$ for $SE(3)$ is defined as following:

$$se(3) = \left\{ \begin{bmatrix} \Omega & \mathbf{v} \\ \mathbf{0}^{1\times 3} & 0 \end{bmatrix} \middle| \Omega \in so(3), \mathbf{v} \in \mathbb{R}^3 \right\}, \tag{3}$$

Given an element $U \in se(3)$ on the tangent space $\mathscr{T}_I SE(3)$ at the identity I of $SE(3)$, the corresponding element $T \in SE(3)$ can be evaluated just by using the exponential map: $\exp : se(3) \to SE(3)$, where $\exp$ in $SE(3)$ is the matrix exponential. The inverse mapping is $\log : SE(3) \to se(3)$, where $\log$ in $SE(3)$ is the principal matrix logarithm. The same mappings hold when restricting to $SO(3)$. Elements of $se(3)$ can be associated with the tangent vector of a curve $A(t) \in SE(3)$, at t, representing the local motion of a rigid body. Elements of this kind are called *twists*, and can be uniquely represented by a 6-dimensional vector $(\boldsymbol{\omega}(t)^\top, \mathbf{v}(t)^\top)^\top$, physically corresponding to the *instantaneous* angular velocity and the *instantaneous* linear velocity of the body, both expressed in the moving body reference frame. The operation $(\cdot)^\vee$ converts a 4×4 twist into the 6 dimensional vector $(\boldsymbol{\omega}(t)^\top, \mathbf{v}(t)^\top)^\top$.

Given a metric specifying properties of the rigid body, [32] show that a geodesic is a locally length-minimizing curve on a manifold, such that, for two configurations $A, B \in SE(3)$:

$$A = \begin{bmatrix} R_A & \mathbf{d}_A \\ \mathbf{0} & 1 \end{bmatrix} \quad B = \begin{bmatrix} R_B & \mathbf{d}_B \\ \mathbf{0} & 1 \end{bmatrix} \tag{4}$$

the geodesic $\Gamma(t)$ is:

$$\Gamma(t)=\begin{bmatrix} R_A \exp(\Omega_0 t) & (\mathbf{d}_B-\mathbf{d}_A)t+\mathbf{d}_A \\ \mathbf{0} & 1 \end{bmatrix} \quad (5)$$

Here $\Omega_0 = \log(R_A^\top R_B)$. The problem to solve in this preliminary part is the following: given a set of Euclidean transformations $T_1, ..., T_n \in SE(3)$, find the principal directions maximizing the variance of the data. This can be obtained by applying the Principal Geodesic Analysis (PGA) introduced for the first time in [12], which is a generalization of PCA when a manifold is considered. The authors define the variance, the subspaces and the projections in a manifold setting. In particular the subspaces, that in PCA were linear, now are *geodesic sub-manifolds*. An extension of the algorithm provided in [12] to $SE(3)$ is straightforward and illustrated in Algorithm 1. Indeed, given the set of body transformations, the centroid $\bar{T}$ is computed, so as to minimize the distance of $\bar{T}$ with all the Ts in the starting set. If the Ts are close enough to each other, it is known that the centroid is unique as stated in [20, 18]. This is the intrinsic mean on the manifold, a generalization to $SE(3)$ is straightforward.

Data: $T_1, ..., T_n \in SE(3)$
Result: Principal directions $\mathbf{e}_i \in \mathscr{T}_\mu SE(3)$ (tangent space of $SE(3)$ at μ) with associated variances $\lambda_i \in \mathbb{R}$
1) Compute $\mu = [\bar{R}\,|\bar{d}]$ with $\bar{R}$ Karcher Mean in $SO(3)$ [20] and $\bar{\mathbf{d}} = 1/n \sum_i \mathbf{d}_i$ on $T_1, ..., T_n$;
2) Compute $\Gamma_{\mu,T_i}(t), t \in [0,1]$ as in eq.(5) with R_A replaced by $\bar{R}$ and R_B replaced by R_i, obtained from $T_i, i = 1, \dots n$ (eq. (1));
3) $\forall T_i$ compute the twist $U_i = \Gamma^{-1}_{\mu,T_i}(t)\dot{\Gamma}_{\mu,T_i}(t)$, $t \in [0,1]$;
4) Compute the vector $(\boldsymbol{\omega}(t)^\top, \mathbf{v}(t)^\top)_i^\top = U_i^\vee$;
5) $S = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{\omega}(t)^\top, \mathbf{v}(t)^\top)_i^\top (\boldsymbol{\omega}(t)^\top, \mathbf{v}(t)^\top)_i$;
6) $\{\lambda_i, \mathbf{e}_i\}$ = eigenvalues and eigenvectors of S;

Algorithm 1: Principal Geodesic Analysis in $SE(3)$

Fact: The twist U_i physically interprets the local motion of a joint, and using its vector representation $(\boldsymbol{\omega}(t)^\top, \mathbf{v}(t)^\top)_i$, we obtain that S, is in $\mathbb{R}^{6\times6}$, and clearly symmetric. Each principal direction $\mathbf{e}_i$, resulting from the PGA algorithm, as an eigenvector of S is in $\mathbb{R}^6$. As Γ_{μ,T_i} is a geodesic, the product $(\Gamma^{-1}_{\mu,T_i}\dot{\Gamma}_{\mu,T_i})$, once applied the $\vee$ transformation, according to the fact that a twist can be uniquely represented by a 6-dimensional vector, specifies the motion between the joint and the Karcher mean $\bar{R}$.

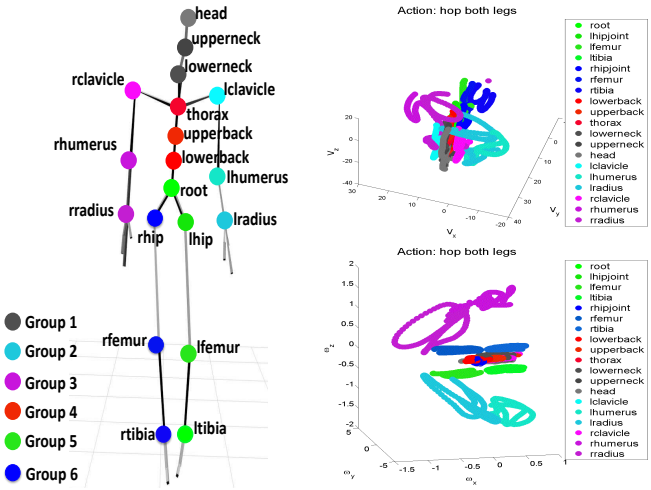

Figure 1: On the left a skeleton with the whole set of joints, groups are highlighted by color. On the right joints motion with respect to $\mathbf{v}$ and $\boldsymbol{\omega}$ highlighting motion similarities within groups (better seen in color).

3. Action Representation Model

In MoCap representation, input data are sequences of joints configurations. Each sequence is about a single subject performing a specific action. Joints are associated with a subject skeleton and are expressed along time as transformation matrices, of the form given in eq. (1), with respect to the global coordinate system. We consider $K = 19$ joints, see Figure 1, left. To properly obtain a representation for each sub-body part we introduce some notation.

Notation In the following we denote j_i an unordered sequence of frames of the action A_i, which we call *sample sequence*. The length of each sample sequence j_i, is denoted by L_{j_i}. Given N_i sample sequences for action A_i, $j_i = 1_i, \dots, N_i$, their length is $L_{1_i}, \dots, L_{N_i}$. Each sample sequence is divided in 6 groups, indexed by m. A feature vector of a number of sample sequences for action A_i is $v^l_{j_i,m}$, where $m = 1, \dots, 6$, $j_i = 1_i, \dots, N_i$, and the superscript l varies on the sequence length.

D_{j_i} denotes the block matrix for the MoCap joints transformations, for each sample sequence j_i:

$$D_{j_i} = \begin{bmatrix} T^1_{j_i,1} & T^1_{j_i,2} & \cdots & T^1_{j_i,K} \\ \vdots & \vdots & \vdots & \vdots \\ T^{L_{j_i}}_{j_i,1} & T^{L_{j_i}}_{j_i,2} & \cdots & T^{L_{j_i}}_{j_i,K} \end{bmatrix}, \quad (6)$$

Here each block $T^l_{j_i,k}$, $k = 1, \dots, K$, is a 4×4 transformation matrix (see eq. (1)) with respect to the world's inertial frame of the sample sequence j_i of action A_i, relative to the k-th joint in frame l.

C_{j_i} denotes the block matrix of all the configurations of a single sample sequence j_i of action A_i, taking into account

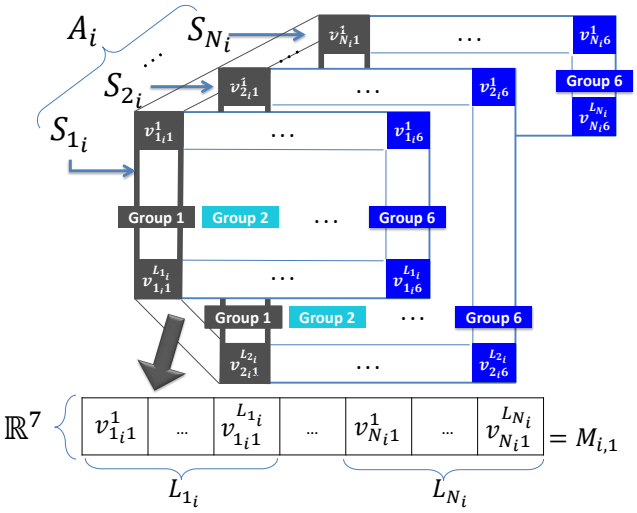

Figure 2: Stack of feature vectors $\mathbf{v}^l_{j_i,m}$ of the first group ($m = 1$) of joints into a $7 \times (L_{1_i} + L_{2_i} + \cdots + L_{N_i})$ matrix.

all the 6 sub-body groups:

$$C_{j_i} = \begin{bmatrix} g^1_{j_i,1} & \cdots & g^1_{j_i,6} \\ \vdots & \vdots & \vdots \\ g^{L_{j_i}}_{j_i,1} & \cdots & g^{L_{j_i}}_{j_i,6} \end{bmatrix}, \tag{7}$$

Here each $g^l_{j_i,m}$ is a block of the form $(T^l_{j_i,a}, \ldots, T^l_{j_i,b})$, of dimension $(4\times4)\cdot h$, with h the number of joints of the m-th sub-body group, for $m = 1, \ldots, 6$ and $1 \leqslant a < b \leqslant K$.

Matrices like C_{j_i} are used to compute the features of sample sequences of action A_i, as shown in Algorithm 2.

Data: N_i sample sequences C_{j_i}, as in eq. (7), of lengths L_{j_i} for action class A_i
Result: Feature vectors of action A_i organized into matrices $\{M_{i,m}\}_{m=1,\ldots,6}$

1. For each block $g^l_{j_i,m}$, of C_{j_i}, compute the first principal direction $\mathbf{e}^l_{j_i,m} \in se(3)$, according to Algorithm 1.
2. Map $\mathbf{e}^l_{j_i,m}$ into a transformation matrix $T^l_{j_i,m} \in SE(3)$, via exponential mapping.
3. Build the feature vector $\mathbf{v}^l_{j_i,m} \in \mathbb{R}^7$, using the rotation angles and the translation obtained from $T^l_{j_i,m}$, and the norm of the instantaneous linear velocity, obtained from $\mathbf{e}^l_{j_i,m}$.

for $m = 1 : 6$ **do**
  $M_{i,m} = \left[\mathbf{v}^1_{1_i,m}, \ldots, \mathbf{v}^{L_{1_i}}_{1_i,m}, \ldots, \mathbf{v}^1_{N_i,m}, \ldots, \mathbf{v}^{L_{N_i}}_{N_i,m}\right]$;
end

Algorithm 2: Features extraction

4. Classification via preferences on DPM

In this section we investigate how to specify an action via a number of behaviors, generated by the body parts involved in the action, and show how to model the action classification problem via the Dirichlet process mixtures. The approach, in this basic formulation, proves that temporal alignment (see e.g. [29]) can be avoided, in so significantly improving the classification process. We introduce first some notation for this section.

Notation: matrix $M_{i,m}, m = 1, \ldots, 6$, $i = 1, \ldots n_A$, n_A the number of actions, collects the sampled sequences of the m-th group of action A_i; a single element of $M_{i,m}$ is a realization $\mathbf{x}_{i_m,k}$ of a stochastic variable $X_{i,m} \in \mathbb{R}^7$, realizations are indicated with lowercase letters. The length of the collected sequences for action A_i is $\sum_{j_i=1}^{N_i} L_{j_i}$. Since only a subset of $M_{i,m}$ is considered for training (see Algorithm 3), we indicate the length of the training data for the i-th action, m-th group, by J_{i_m}. The training set for action A_i, group m, is $\mathscr{X}^\circ_{i,m} = \{(\mathbf{x}_{i_m,k}, y_{i_m,k}) | k = 1, \ldots, J_{i_m}\}$, $i_m = 1, \ldots, n_A$. A query $Q^\star_m$, $m = 1, \ldots, 6$, is a set of variables $\{\mathbf{x}_{i_m,j_1}, \ldots, \mathbf{x}_{i_m,j_q}\}$, $\mathbf{x}_u \in \mathbb{R}^7$, i.e. any permutation of a set of elements, sampled from a possibly temporally ordered MoCap sequence. It is, thus, intended that $Q^\star_m$ is the result of the transformation of the joints of an observed action via PGA, and it is related to a group m. The classification problem is to classify, for each $m = 1, \ldots 6$, the query $Q^\star_m$, and issue a label $y \in \{\ell_1, \ldots, \ell_{n_A}\}$ for the observed action. In the following we shall omit the subscript m in $\mathbf{x}_{i_m,k}$ and $y_{i_m,k}$.

Recall that the feature vectors are obtained from the principal directions of a group of joints whose rigid motions are referred to a global frame. Therefore within the set of observations for the same group the response vectors are considered an exchangeable sequence, and the ordering is irrelevant. Given a training set $\mathscr{X} = \cup_i \cup_m \mathscr{X}^\circ_{i,m}$, if the parameters are known, hence $p(\Theta|\mathscr{X})$ can be estimated, then $Q^\star_m$, $m = 1, \ldots, 6$, can be classified basing on the predictive densities:

$$P_i(y^\star_{i,k} = y|\mathscr{X}, Q^\star_m) = \int_{\mathscr{D}} p(y^\star_{i,k} = y|\mathscr{X}, Q^\star_m, \Theta) p(\Theta|\mathscr{X}, Q^\star_m) d\Theta \tag{8}$$

With $\mathscr{D}$ the domain and Θ the vector of all parameters in the model. Then using the loss function based on the percentage of correctly classified, the label assigned to each group m is estimated by the maximum a posteriori MAP:

$$\hat{y} = \arg\max_y \{p(y^\star_{i,k} = y|\mathscr{X}, Q^\star_m)\} \tag{9}$$

The basic steps for classification and prediction are illustrated in Algorithms 3 and 4. The probability model that we consider for the classification problem is the popular Dirichlet process (DP) mixtures (DPM)[10, 6]. A DP places a distribution on the space of distributions, generating a distribution on the countable set of mixtures, hence we consider a set of DPMs, one for each group m, of each action A_i. Consider the multivariate normal $\mathcal{N}_7(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^7$, $\boldsymbol{\Sigma} \in \mathbb{R}^{7\times 7}$ and $\boldsymbol{\theta}_{i,j} = (\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}), j < \infty$:

$$\begin{array}{l} \mathbf{x}_{i,k}|\boldsymbol{\theta}_{i,j} \sim \mathcal{N}_7(\mathbf{x}_{i,k}|\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}) \\ \boldsymbol{\theta}_{i,j}|G_m \sim G_m \\ G_m \sim DP(\alpha H) \end{array} \tag{10}$$

Here we are assuming that observations are i.i.d sampled from a parametric family, namely a multivariate Gaussian distribution, with parameters $\boldsymbol{\theta}_{i,k}$, which are in turns independently sampled from an unknown distribution G_m on which is placed a Dirichlet process $DP(\alpha H)$. Where α is the concentration parameter affecting the number of clusters that will be generated, and H is the base distribution. Namely, for a subset of $\mathscr{X}$, $H(A) = E[G_m(A)]$, and typically H is taken to be the conjugate prior of the observation distribution. Here we follow the conjugate approach for the multivariate normal, by choosing:

$$\begin{array}{lcl} (\boldsymbol{\Sigma}_{i,j}|\beta, W) & \sim & \mathscr{W}(\beta, \beta W^{-1}) \\ (\boldsymbol{\mu}_{i,j}|\boldsymbol{\Sigma}_{i,j}, \boldsymbol{\nu}, \rho) & \sim & \mathcal{N}(\boldsymbol{\nu}, (\boldsymbol{\rho}\boldsymbol{\Sigma}_{i,j})^{-1})) \\ (\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}) & \sim & \mathcal{N}_{\mathscr{W}}(\boldsymbol{\nu}, \boldsymbol{\rho}, \beta, \beta W) \end{array} \tag{11}$$

Here $\mathscr{W}$ is the Wishart distribution, with $\beta > 7$ DOF, 7 the dimension of $\mathcal{N}_7(\cdot)$. $\mathcal{N}_{\mathscr{W}}$ is the normal Wishart joint prior distribution with $\boldsymbol{\nu}, \boldsymbol{\rho}, \beta, \beta W$ common to all mixture components of the group m. In turn the priors for $\boldsymbol{\nu}$ and $\boldsymbol{\rho}$ are Gaussian and Gamma, while for W and β the priors are the Wishart and Gamma (see [16, 27] for further details).

The unknown distribution is evaluated at observation points and, according to its discreteness, generates clusters of observations. Namely, in any sample $\boldsymbol{\theta}_{i,1}, \ldots, \boldsymbol{\theta}_{i,j}$ from G_m there is a positive probability of identical values (see [9, 10]). Then each sample can either be assigned to an existing partition or it can generate a new one. This is regulated by the probabilities $n_h/(\alpha+n-1)$ and $\alpha/(\alpha+n-1)$, which induce the Chinese restaurant process (CRP), and the mixing proportion probabilities $\boldsymbol{\pi}_{i,j}$. Where n_h is the number of elements of the cluster to which the repeated sample $\boldsymbol{\theta}_{i,h}$ would belong to.

Inference of the parameters and hyperparameters is obtained for each group by Gibbs sampling and updating them from their posterior distribution as specified above, using the steps for conjugate prior as in [22] and adopting the clever solutions indicated in [27]. Many approaches have highlighted the need to investigate the dependences among data in different groups when these are generated by DPMs,

Data: $\mathscr{X}_{i,m}, i{=}1,\ldots,n_A, m{=}1,\ldots,6$
Result: Parameters
$\Theta_{i,m}, K_{i,m}, i{=}1,\ldots,n_A, m{=}1,\ldots,6$
for $i = 1 : n_A$ **do**
  for $m = 1 : 6$ **do**
    $\mathscr{X}^{\circ}_{i,m}$:= draw sample training data from $\mathscr{X}_{i,m}$;
    $Test_{i,m} := \mathscr{X}_{i,m} \backslash \mathscr{X}^{\circ}_{i,m}$;
    $\boldsymbol{\theta}_{j_i,j_m}$:= estimate parameters using eq.(11), via DPM;
    Fix $K_{i,m}$ as new clusters approach zero;
    $\Theta_{i,m} = \{\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_{K_{i,m}}, \boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_{K_{i,m}}, \boldsymbol{\pi}_1, \ldots \boldsymbol{\pi}_{K_{i,m}}\}$;
  end
end

Algorithm 3: Basic steps in parameters estimation

since the work of [8]. Where, in particular, the problem of how to determine clusters of data in the presence of partial exchangeability and unknown partition of the observations, has been addressed. A solution has been indicated in [28] via the hierarchical DPM (HDPM), which can discover dependencies, generating shared clusters with different weights but same locations.

In the representation we propose, considering the domain of the sub-body part features, two subgroups might take values in space regions that intersect. Despite this the range are usually different and also the observations come separated at the source and the groups are known. Therefore, we combine the groups, in terms of the behaviors that are generated by the DPM for each of them, and use the MAP on the combined groups. To this end we define a preference matrix F of size $n_A \times 6$, with n_A the number of action classes considered. The stochastic matrix F, which provides the optimal combination for the groups, is a matrix of multinomial variables, evaluated according to a success matrix S. Each row of F represents the experiment assigning a success to the group m, which provides the best contribution to characterize the action. This is assessed by assigning a success to the group that has higher concentration parameter, since this is sensible to the number of behaviors, which implies that the group undergoes several changes during the action execution, hence the involved sub-part characterizes the action. The successes recorded for the multinomial at S are the values of the concentration parameter α estimated for the DPM of the group. The parameters of F are estimated at the final step of the Gibbs sampling and kept common to all the groups estimation. An initialization of S is provided assigning a success to the group/groups that are considered the more active ones in the

action execution, according to a rule of thumb. The com-

Data: $Q^{\star}_m, \Theta_{i,m}, K_{i,m}, S_{i,m}$, $i = 1, \ldots, n_A, m = 1, \ldots, 6$
Result: Matrix $\mathscr{M}$ for $\cup_{m=1}^{6} Q^{\star}_m$
for $i = 1 : n_A$ **do**
  for $m = 1 : 6$ **do**
    $p(Q^{\star}_m|\Theta_{i,m}, K_{i,m}) \propto$
    $\sum_{j=1}^{K_{i,m}} \mathscr{N}_7(Q^{\star}_m|\boldsymbol{\theta}_{i,j})\boldsymbol{\pi}_j$;
      with $(\boldsymbol{\theta}_{i,j}, \boldsymbol{\pi}_{i,j}) \in \Theta_{i,m}$;
    $Z_{i,m} := p(Q^{\star}_m|\Theta_{i,m}, K_{i,m})$;
  end
end
Final step
Compute the new mixture weights F according to eq.(12);
$\mathscr{M} = F^{\top} Z$;

Algorithm 4: Basic steps of prediction

putation of F is carried at the last step of Gibbs sampling. Considering that each group is evaluated in turn, for each action i, F_i is the i-th row of F corresponding to the current evaluated action. Let $\kappa_{i,m}$ be the prior assigned to the Dirichlet distribution for the group m and t the final step of Gibbs sampling:

$$F^{(t)}_{i,m} = \frac{S^{(t)}_{i,m} + \kappa_{i,m}}{n_A + \sum_{m=1}^{6} S^{(t)}_{i,m}} \tag{12}$$

Then the new mixture is obtained simply as $F^{\top} Z$, where Z is the matrix of the DPM distributions computed for each group m. We can note that the final mixture is still a mixture combining the DPM models for each group, weighting the groups in a way sensible to the number of behaviors elicited by the DPM model. Without a non-parametric approach this last mixing, which so to say meta-evaluates the estimation, would have not been possible.

5. Experiments and Results

In this section we report experimental results on the performance of the proposed method for MoCap action recognition. The goal of the experiments is to verify the accuracy of the prediction of a new observed action.

Data We consider 11 types of "cut actions" (i.e. a single type of action per sequence) obtained from HDM05 [21], where each cut action is performed by 4 different subjects, and similar types of actions from CMU [1]. Results from [1] are not reported, though almost the same, being the data

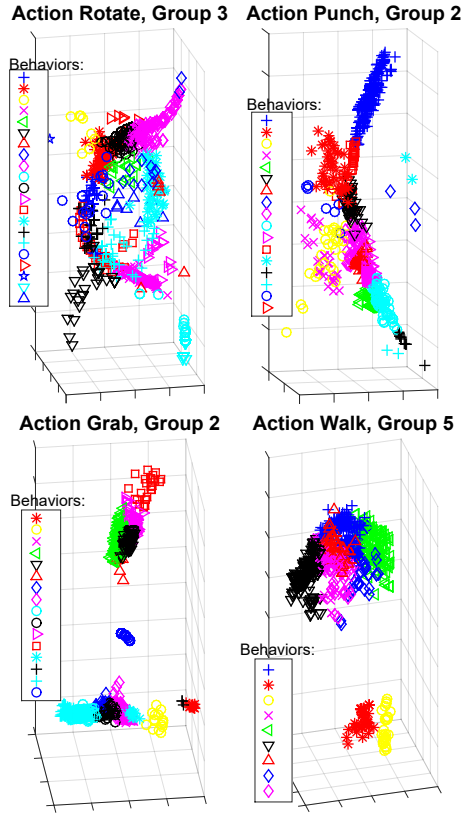

Figure 3: Behaviors clustering for 4 sub-body parts of 4 different actions.

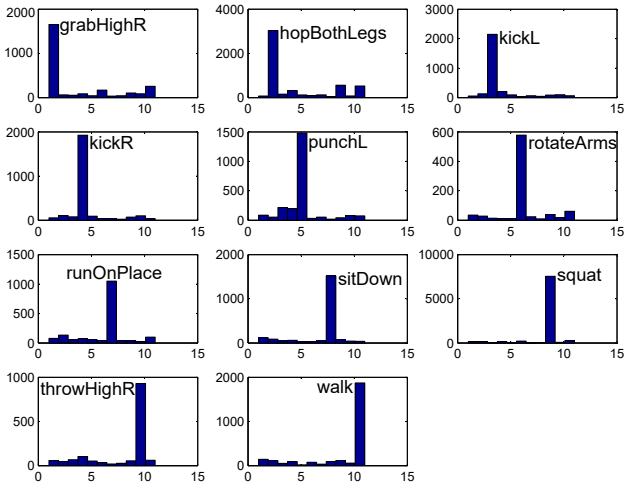

Figure 4: Histograms of MAP response for each Action Category

noiseless. The actions considered from [21] are: grab an object from high with right arm (3401 frames), hop with both legs (5941 frames), kick with left leg (3828 frames), kick with right leg (3374 frames), punch with left arm (3144

Action Class	#Clusters Group 1	#Clusters Group 2	#Clusters Group 3	#Clusters Group 4	#Clusters Group 5	#Clusters Group 6
Kick with Left Leg	10	17	14	10	**26**	15
Throw with Left Arm	13	**23**	15	12	19	18
Squat	12	13	13	3	11	13
Walk	8	11	10	4	9	8

Table 1: Number of clusters generated for each group of joints for 4 different categories of actions

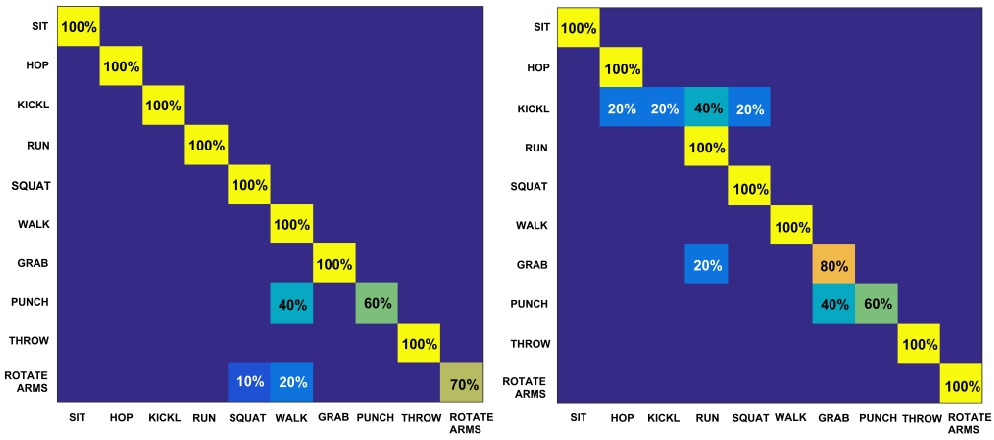

Figure 5: Confusion matrices for comparing the PGA-DPM algorithm, on the left, with the one presented in [14] by Gong and Medioni, on the right.

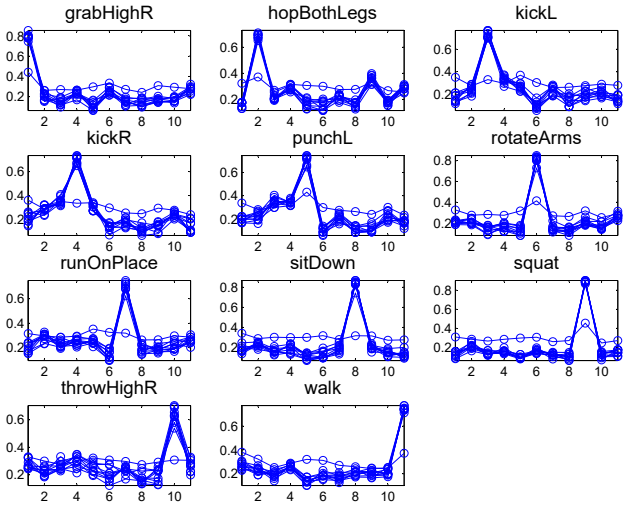

Figure 6: MAP evaluation with repeated random samples from test data, for each Action Category

frames), rotate both arms backward (1632 frames), run on place (139440 frames), sit down on chair (2884 frames), squat (9519 frames), throw an object with right arm (2254 frames), walk (3470 frames). We have also considered the datasets [2, 3, 4, 5], and adapted it to our full 3D model. Despite these datasets are noisier than CMU and HDM05, results are comparable but not reported for lack of space.

Method All data available are structured according to the description provided in Section 3, then they are transformed to obtain the PGA features according to the description provided in Section 2. We have trained the DPM model as follows. For each action we consider 800 data for training. From this set we then define the training set for each group by randomly sampling from the chosen training set. All the remaining frames are considered for test. Running the Gibbs sampler we obtain a model for each group of each action and we store it into a data structure. We distinguish between a set of frames, randomly chosen from a sequence of frames, in which data are ordered according to the action evolution.

Now, given a set of frames (or an action sequence) from the data test, we first estimate the probability of each group according to the parameters of the model and the mixture components and then we combine the groups using the estimated weight matrix F, eq. (12). The resulting classification is obtained by MAP estimation, eq. (9). Estimation of either a set or a sequence of actions takes less than one sec. of computation time. Similarly geometric transformations and features computation are on the order of 10^2 sec. On

Approach	Total Accuracy
PGA-DPM	93.86%
DMW	85.78%

Table 2: Total Accuracy for the PGA-DPM based method and for DMW[14]

the other hand the computational cost for learning is quite high, of the order of 10^6 sec.

Experiments We have conducted the following experiments. In the first experiment we have tested all the test data and verified the MAP on the whole set, this is illustrated in Figure 4. Each panel, in the figure, shows the histogram of the classification on the whole test set. We can note that the maximum is always correctly assigned. In the second experiment, given that the number of test data is N, we have randomly sampled from them $N/10 + k$, $k > 10$ data and the results are reported in Figure 6.

Finally we have extracted actions as sequences from the test data and the classification results are reported in the confusion matrix in Figure 5, where the results have also been compared with [15].

Comparisons We have chosen the algorithm of Dynamic Manifold Warping [15, 14]. DMW is basically an instance-based learning in which the action sequences are represented as structured time series. The authors, in [15], first temporally align the testing sequence with all the training labeled sequences. They then extract for each aligned sequences' frames a similarity measure between the testing sequence and the temporally aligned training sequences, and the action performed in the testing sequence is labeled with the label of the training sequence from which the testing sequence has minimum distance. In our approach, instead, we learn a model so as to estimate the most representative behaviors made by each of the groups of joints, not considering structured sequences along time, but considering, instead, each feature conditionally independent on the other ones. Therefore, while DMW depends on the sequences considered and for each new input sequence has to compare it with all the labeled training sequences, our algorithm has a learning process so that the testing process is immediate and the accuracy in recognition increases with the number of features considered in the DPM process, following the "rich get richer" fashion, typical of the DPMs. It is worth mentioning that in order to evaluate the DMW accuracy, we have implemented a version of DMW with a choice of parameters and methods that are hidden in [15].

In order to compare our algorithm with DMW, we have considered 10 configuration sequences of PGA-based features for each action category group. We have used the term configuration sequence, since in our model we do not have ordered data, but instead features that are exchangeable. The tests have been made on 10 actions. In this case, the MAP estimate for our algorithm is computed for each single query frame of a configuration sequence, and the accuracy for each query sequence is evaluated as the percentage of correctly recognized query frames in the query sequence over the total number of frames of that sequence. For DMW, instead, the accuracy is simply the number of sequences correctly recognized, over the total number of sequences. In Figure 5, it is possible to see the confusion matrix for our approach and for DMW. In Table 2, it is shown for the two approaches the accuracy computed as the total number of recognized query frames over the total number of considered sequences.

Evaluation Table 1 shows the number of clusters estimated by the PGA-DPM (as explained in Section 4) for each of the sub-body group of joints for 4 different types of actions. Note that in the kick and throw actions, a large number of clusters is estimated for the most representative groups of joints (i.e. the left leg and the left arm, respectively). For the squat and the walk actions, instead, excluding the joints of the torso (group 4), all the sub-body groups are involved in the motions, and therefore a more distributed number of clusters is estimated. Furthermore, in Figure 3 it is possible to visualize some of the generated clusters for an arbitrary sub-body group in 4 different actions categories: kick with left leg, rotate arms, punch with left arm, walk.

6. Conclusions

We have presented a novel approach to the human action recognition problem, by considering a new MoCap feature representation, which has been verified to be suitable for developing a non-parametric Bayesian method for classification, via the DPM. In particular, we have combined the skeleton joints into groups and reduced their dimensionality by means of PGA, so as to maintain a solid information on motion. Assuming features to be conditionally independent, for each group, given a specific prior, we have applied DPM to generate the most representative behaviors for each group of joints and each action category so as to perform classification. Our approach proves that a time-ordered representation for MoCap sequences is not needed and indeed, as shown in Section 5, performances are good and our approach outperforms exactly time-alignment based approaches as [14]. Basing on these promising results we are now investigating more complex actions, in particular the collaborative ones, in which two different subjects must pass objects between them, and carry objects together.

Acknowledgments
Supported by the EU FP7 TRADR (609763) and the EU H2020 SecondHands (643950) projects.

References

[1] http://mocap.cs.cmu.edu. 6

[2] http://http://www.cs.ucf.edu/~oreifej/HON4D.html. 7

[3] http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/. 7

[4] http://www.micc.unifi.it/vim/datasets/3dactions/. 7

[5] http://dipersec.kingston.ac.uk/G3D/. 7

[6] C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, 1974. 5

[7] V. Bloom, D. Makris, and V. Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *CVPRW*, 2012. 2

[8] D. M. Cifarelli and E. Regazzini. Distribution functions of means of a dirichlet process. *The Annals of Statistics*, 1990. 5

[9] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 1995. 5

[10] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, 1973. 5

[11] F. Flaherty and M. do Carmo. *Riemannian Geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, 2013. 2

[12] P. Fletcher, C. Lu, S. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *MI, IEEE Transactions on*, 2004. 3

[13] S. Fothergill, H. M. Mentis, S. Nowozin, and P. Kohli. Instructing people for training gestural interactive systems. ACM, 2012. 2

[14] D. Gong and G. Medioni. Dynamic manifold warping for view invariant action recognition. In *ICCV, 2011 IEEE International Conference on*. IEEE, 2011. 1, 2, 7, 8

[15] D. Gong, G. Medioni, and X. Zhao. Structured time series analysis for human action segmentation and recognition. In *IEEE Transactions on PAMI*. IEEE Computer Society, 2014. 2, 8

[16] D. Görür and C. E. Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *J. Comput. Sci. Technol.*, 2010. 5

[17] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In *ECCV 2014*. Springer, 2014. 1

[18] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 1977. 3

[19] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *ECCV, 2006*. Springer, 2006. 1

[20] J. Manton. A globally convergent numerical algorithm for computing the centre of mass on compact lie groups. In *ICARCV*, 2004. 3

[21] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation Mocap Database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007. 6

[22] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 2000. 5

[23] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In *CVPRW, 2012 IEEE Computer Society Conference on*, 2012. 1, 2

[24] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR, 2013 IEEE Conference on*. IEEE, 2013. 2

[25] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *CVPRW*, 2013. 2

[26] M. Spivak. Differential geometry, volume 1–5. *Publish or Perish, Berkeley*, 1975. 2

[27] E. B. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, MIT, 2006. 5

[28] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006. 5

[29] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR, 2014 IEEE Conference on*, 2014. 1, 2, 4

[30] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *PAMI, IEEE Transactions on*, 2014. 1, 2

[31] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *CVPRW, 2012 IEEE Computer Society Conference on*. IEEE, 2012. 2

[32] M. Zefran, V. Kumar, and C. Croke. Choice of riemannian metrics for rigid body kinematics. In *ASME 24th Biennial Mechanisms Conference*, 1996. 2