Bayesian Image Based 3D Pose Estimation

Marta Sanzari(✉), Valsamis Ntouskos(✉), and Fiora Pirri(✉)

ALCOR Lab, DIAG, Sapienza University of Rome, Rome, Italy
{sanzari,ntouskos,pirri}@diag.uniroma1.it

Abstract. We introduce a 3D human pose estimation method from single image, based on a hierarchical Bayesian non-parametric model. The proposed model relies on a representation of the idiosyncratic motion of human body parts, which is captured by a subdivision of the human skeleton joints into groups. A dictionary of motion snapshots for each group is generated. The hierarchy ensures to integrate the visual features within the pose dictionary. Given a query image, the learned dictionary is used to estimate the likelihood of the group pose based on its visual features. The full-body pose is reconstructed taking into account the consistency of the connected group poses. The results show that the proposed approach is able to accurately reconstruct the 3D pose of previously unseen subjects.

Keywords: Human pose estimation · Hierarchical non-parametric Bayes

1 Introduction

Human pose estimation from images has been considered since the early days of computer vision and many approaches have been proposed to face this quite challenging problem. A large part of the literature has concentrated on identifying a 2D description of the pose mainly by trying to estimate the positions of the human joints in the images. Recently, attention has been shifted to the problem of recovering the full 3D pose of a subject either from a single frame or from a video sequence. Despite this is an ill-posed problem due to the ambiguities emerging by the projection operation, the constraints induced by both human motion kinematics and dynamics have facilitated the recovery of some accurate 3D human pose estimation.

In this work we approach the problem of 3D pose estimation from a single image building a hierarchical framework based on Bayesian non-parametric estimation. A schema of the framework is shown in (Fig. 3). Following the schema flow, we divide the human body into different parts and we study the idiosyncratic motion behavior of each part independently from the others. In this way

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46484-8_34) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part VIII, LNCS 9912, pp. 566–582, 2016.
DOI: 10.1007/978-3-319-46484-8_34

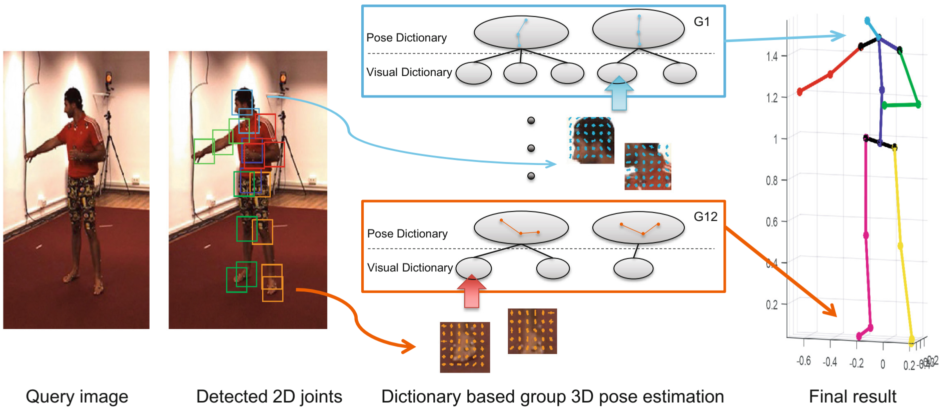

Fig. 1. Method overview; 3D pose estimation given a query image.

we learn the principal motion modes of each part. Each body part is specified by a group of joints, and its motion is represented by pose features obtained by the principal motion direction on the $SE(3)$ manifold with respect to a reference pose. As a natural reference pose we consider the "Vitruvian man" pose presented in Fig. 2 together with the selected groups.

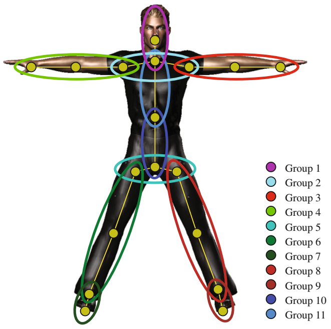

Fig. 2. "Vitruvian" pose with defined groups.

The visual features for each group are the PHOG features of [1], which are computed using the state-of-the-art approach of [2]. Assuming a correspondence between the visual and pose features both the space of visual features and pose features are partitioned, in such a way that from the visual features it is possible to accede to the non observed pose features. These nested partitions are built up for each group with a hierarchical non-parametric Bayesian model, designed purposefully to deal with the inverse projection problem, from 2D to 3D. Indeed, the goal is to recover the unknown human poses just from the available visual features, since visual features are the only available observations.

The hierarchical model is based on two nested countably infinite mixtures of normal distributions. The first level builds a dictionary of 3D human poses by considering various examples of 3D human poses taken from a large number of motion sequences, while the second level takes into account the corresponding images obtained from a number of view points. Indeed, the dictionary is built by partitioning the space of 3D poses with a Dirichlet process mixture model (DPM). The partition is defined on the space of poses specified by the principal motion directions on the $SE(3)$ manifold. The nested part of the model builds the visual dictionary on top of the pose dictionary, and it is also based on Dirichlet process mixture models. Here the mixture processes the PHOG [1] features extracted from a window centered at the 2D position of each joint in the given image.

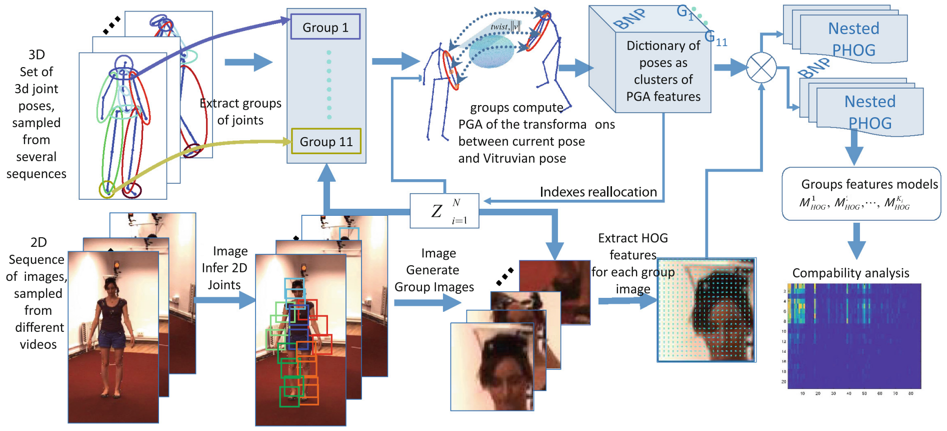
3D Set of 3d joint poses, sampled from several sequences
Extract groups of joints
Group 1
Group 11
twist, ||v||
groups compute PGA of the transforma ons between current pose and Vitruvian pose
BNP
G1
G11
Dictionary of poses as clusters of PGA features
Nested PHOG
BNP
Nested PHOG
Groups features models
$M^1_{HOG}, M^i_{HOG}, \cdots, M^{K_i}_{HOG}$
Indexes reallocation
$Z^{N}_{i=1}$
2D Sequence of images, sampled from different videos
Image Infer 2D Joints
Image Generate Group Images
Extract HOG features for each group image
Compability analysis

Fig. 3. Schematic representation of the proposed hierarchical model.

Based on the learned dictionary 3D pose estimation is performed as follows (Fig. 1). Given a query image we extract the 2D positions of the joint in the image using a state-of-the-art approach [2] and compute the corresponding PHOG features for each group. From these features we infer the most likely cluster of the visual dictionary, which in turns indicates the cluster of 3D poses with the highest probability for the given group. The final 3D pose is reconstructed by assembling together the most representative poses of the selected clusters for each group. Clusters are selected considering also the compatibility between the group poses.

In the following, Sect. 2 discusses related work and Sect. 3 the structure of the training and testing data, and preliminaries. Section 4 presents the architecture of the proposed model and how pose estimation is performed. In Sect. 5 we present the results obtained with our method in comparison with state-of-the-art 3D pose estimation approaches. Finally, Sect. 6 discusses conclusions and future work.

2 Related Work

Human pose estimation (HPE) has been extensively studied during the years by considering videos, 2D images and depth data, [3–5]. There exist several open problems; among them we mention variations in human appearance, clothing and background, arbitrary camera view-point, self-occlusions and obstructed visibility, ambiguities and inconsistency in the estimated poses.

Different features can be chosen to describe the different types of data. Focusing on 2D input data, some works assume the 2D body joints locations already given [6], while others extract features from silhouettes such as HOG [7], PHOG [1], SIFT [8] and shape context [9], or dense trajectories [10].

In detail, concerning 3D HPE from videos, very recently [10] introduced a spatio-temporal matching (STM) among 3D Motion Capture (MoCap) data and 2D feature trajectories providing the estimated camera view-point and a selected subset of tracked trajectories. In our approach, instead, as in [11,12], body parts in 2D are detected by using the algorithm introduced in [2].

In the last years many works have approached the estimation of the poses via deep learning as in [13–16]. In Zhou et al. [17] a convolutional neural network is used to estimate the 2D joint locations in the image. 3D pose sequences are then estimated via an EM algorithm over the entire video by considering a sparse model of 3D human pose in input where each 3D body pose is represented by a linear combination of a predefined basis of poses. Wang et al. [12] propose an overcomplete dictionary of poses learned from 3D human poses and HPE is managed by minimizing an L_1 norm error between the projection of the 3D pose and the corresponding 2D detection, optimizing via alternating direction method. In [18], body part detectors provide proposals for the location of 2D pose of visible limbs. The 2D pose is then refined via non-parametric belief propagation and the corresponding 3D pose is estimated by learning the parameters of a mixture of experts model.

In [19] a relevance vector machine is proposed to learn a reconstruction function that is a linear combination over a set of basis functions. The authors extract shape descriptors from a set of 2D images and the corresponding 3D poses. [20] store a set of different images and full body poses, both in 2D, together with the corresponding viewpoint. A test image is directly matched with all the training images via the shape context matching procedure. The 3D positions are then estimated via the Taylor's approach [21]. Differently from ours, their methods is instance-based, which is not feasible for a real-time application, without also the possibility of generalizing over the training images.

Assuming that joint positions are already given in 2D with the corresponding image, [6] propose to learn pose-dependent joint angle limits from a MoCap dataset, to form a prior for estimating the 3D poses, together with the camera parameters. A tracking-by-detection technique is used in [22] to collect a small number of consecutive video frames. A novel class of descriptors, called tracklets, is defined and 3D poses are recovered from them. In [23], human pose is estimated via a non-parametric Bayesian network and structure learning, considering the dependencies of body parts. In our approach, instead, nested non-parametric clustering is considered to find relations among the appearance and the 3D pose of each body part. As in [23], our approach is able to generalize over the observed data so as to generate new poses never seen before.

In [24], besides the construction of a large dataset, a benchmark among various HPE approaches is performed. [25] use boolean relationships between body components, called posebits, for training an SVM for retrieving the 3D body pose. Finally, [26] consider annotated 2D images and MoCap data as independent input data to first obtain an initial pose model which is then refined iteratively.

3 Description of Input Data

Human 3.6 M Dataset. The dataset we consider for the development of our HPE algorithm is Human 3.6 M [24], which includes about 3.6 million video frames with associated labelled joints and poses of different human subjects performing actions. Relevant for us are the motion capture (MoCap) data (provided as joints rotations and translations) acquired with the Vicon MoCap System; data of 11 subjects performing 15 different actions are available. The 3D joint poses are provided as transformation matrices evaluated with respect to a fixed world origin as described in the next subsection.

Additionally, we consider the corresponding video frames captured from high resolution RGB cameras from 4 different viewpoints. This is done to ensure that we take in consideration a sufficiently varied set of poses captured from different view points. We consider the 4 views of each pose as distinct instances. Furthermore, we are given also the positions of the MoCap skeleton mapped into the image domain. This is used for the 2D joints inference in images, as explained in the following. As in [17], we use 5 subjects (S1, S5, S6, S7, S8) for the training stages, and 2 subjects (S9, S11) for testing. Moreover, we consider only 18 out of the entire set of 32 3D joints by excluding joints corresponding to fingers and toes and by merging together joints corresponding to the same 3D position in order to avoid redundancy in the data. Therefore, for each video frame we have the association among the image, the 3D joint poses, and the 2D joints mapped in the image.

PGA-Based Features. We now describe the basic principles used for extracting features representing the pose of each group. A MoCap sequence amounts to the poses of a subject at regular time instances. At each time instant the pose of the subject is represented by a given configuration of its joints. In detail, a skeleton $\mathcal{J}$ is specified by 18 joints, where the first one is the index of the root joint. Each joint has a single parent joint, except from the root joint. The configuration of the i-th joint is represented by a homogeneous transformation matrix $T_i \in SE(3)$, a *Lie Group* with identity element defined by the 4 × 4 identity matrix. By defining a proper metric the Lie Group is a Riemannian manifold, on which we can define (via the exponential mapping) the notion of geodesic between two elements on the manifold (see [27–29]), which is locally the shortest path that connects two group elements. Henceforth each joint is considered as a rigid body moving in space with respect to some coordinate system. Note that this coordinate system may change according to the MoCap system used for acquiring the data.

We breakdown the skeleton into 11 sub-body groups G_s, with $s = 1, \dots, 11$. Each group contains M_s joints and is defined as $G_s = \{J_{\psi(1)}, \dots, J_{\psi(M_s)}\} \subseteq \mathcal{J}$, with $\psi(\cdot)$ providing the relation of the group joint indices with respect to the skeleton indexes. All joints belonging to a group have a parent within the same group, except the root of the group, which is included in at least one other group, whenever it is not the root of the entire skeleton, this proviso is required by the reconstruction of the full-body pose (Algorithm 2).

Breaking down the skeleton into groups is motivated by the idiosyncratic motion of body parts, and to appraise this fact we use the Da Vinci's Vitruvian pose as the reference skeleton configuration, adapting an idea of [30]. The Vitruvian pose and the joint groups considered here are shown in Fig. 2. Now, given

Table 1. Average geodesic distance between the Karcher mean and the rotations of each joints for each group over the whole dataset.

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	G_{11}
J_1	1.102	1.152	1.152	1.149	1.144	1.143	1.108	1.145	1.106	1.110	1.141
J_2	1.102	1.521	1.521	1.521	1.524	1.518	1.108	1.535	1.106	1.110	1.510
J_3	-	1.520	1.519	1.519	1.540	1.521	-	1.530	-	-	1.519

a pose, we find the transformation between the current pose configuration and the Vitruvian pose, for each group G_s, $s = 1, \dots, 11$. Then, the pose feature set for each group is obtained from the principal direction, computed via *Principal Geodesic Analysis* [31] from these transformations.

More specifically, for each G_s, $s = 1, \dots, 11$ the transformation matrices mapping the joints from a current arbitrary pose to the Vitruvian pose are computed, taking into account the dependencies from the parent pose. We compute the Karcher mean [32] μ of the group transformations, following the algorithm of Afsari [33]. In particular, regarding rotation averaging, the center of mass should be within a geodesic distance no larger than $\pi/2$ in order to be unique, and thus well defined [33–35]. Table 1 shows the average geodesic distance between the intrinsic mean and the rotations of the individual joints for each group over the whole dataset, suggesting that the Karcher mean computation is well defined for this particular choice of groups.

Hence we compute the tangent space of SE(3) at μ and select the principal direction. This direction is the one that best interprets the variability of the motion that the group of joints performs in order to return to the configuration of joints of that sub-body group, in the Vitruvian rest pose. The actual computation of the principal direction in SE(3) is given in [36], and for the transformation considered here the whole computation is resumed in Algorithm 1.

2D Joints Estimation from Monocular Images. In both learning and testing stages we extract PHOG visual features for each considered group. For this purpose, given an image sampled from a video of the dataset in Human 3.6 M, the first step is the estimation of the 2D joints together with suitable surrounding boxes in the image domain.

In detail, since we have considered the 3D skeleton subdivided into 11 groups we recover 11 boxes (or windows), one for each imaged group. From each of these boxes we extract the most suitable image descriptors for our purpose, that are the Pyramid Histogram of Oriented Gradients (PHOG) [1,7]. We have decided to consider a pyramid with levels equal to 0 and 1 and 8 bins spanning an angle of 360°, for each joint in the group, this choice leads to feature vectors of size $m, m \in \{16, 24, 32\}$.

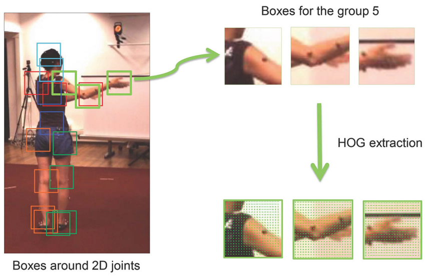

Fig. 4. **Left:** 2D joints estimation using [2]; **Right:** HOG descriptor extraction for a group of joints.

```
Data: The pose of the group $G_s$ given by the corresponding set of
      homogeneous transformations $\{T_{\psi(1)}, \ldots, T_{\psi(M_s)}\}$; the Vitruvian
      joints configuration $\{T^V_{\psi(1)}, \ldots, T^V_{\psi(M_s)}\}$.
Result: Feature vector for the pose of the group $G_s$
  1. Move the root of $G_s$ to the root of the corresponding group in
     Vitruvian pose.
  2. Compute the “disparity” between each joint current pose and the
     Vitruvian pose as $\hat{G}_s = \{\hat{T}_{\psi(1)}, \ldots, \hat{T}_{\psi(M_s)}\}$, taking into account the
     dependency of each joint pose from its parent pose.
  3. Compute the Karcher mean as in [33], extending it to translation.
  4. Compute the variance $S$ as in [31], but using the twist
     $\mathbf{u}^\vee = (\omega^\top, \mathbf{v}^\top)^\top$, obtained from the Lie algebra of the given
     transformations, to extend the PGA to SE(3), with $\omega$ and $\mathbf{v}$ the
     instantaneous angular and linear velocities, as in [36].
  5. Compute the eigenvector and eigenvalues of $S$ and return the first
     principal direction in the Lie algebra $se(3)$.
  6. Build the feature vector in $\mathbb{R}^7$ using the instantaneous angular and
     linear velocities from the principal direction, forming a twist, together
     with the norm of the instantaneous linear velocity [36].
```

Algorithm 1. Feature extraction for the pose of a group G_s of joints

The estimation of the 2D joints from images is performed using the state-of-the-art approach [2]. This approach is particularly suitable for the estimation of the sought-after boxes surrounding joints of human body. We train a model using the algorithm described in [2] using images sampled from the videos in the Human 3.6 M dataset. In particular, we used 61750 images for training taken by the 5 different subjects (S1, S5, S6, S7, S8) performing all the actions, provided together with the 2D joints positions. We used 24700 images for testing taken from the remaining subjects (S9, S11) performing the same actions. From the boxes obtained we consider the central points being the 2D joints. Note that we know the ordering of the parts and so of the joints. Figure 4 shows the result of the boxes extraction for two different testing images and the process of PHOG extraction from an image of a group when the PHOG level is set to 0.

4 Features to Poses Mapping: A Hierarchical Model

In this section we present the hierarchical model connecting 3D poses and visual features, which make it possible to infer a human pose from the visual features. The hierarchical model takes care of the main aspects of this inference process. First of all it generates a dictionary of poses, for each group. The dictionary collects poses in clusters, where the similarity within a cluster is defined according to the parameters of the underlying distribution. In particular, the dictionary for the poses is a list of indexes specifying for each pose the set of poses sharing the same partition block – or the same parameters. Because a set of similar poses admits several views, the visual features indexed in the same partition generate a mixture of features too. Finally, a principle of compatibility amid clusters of different groups is defined.

In this section we consider $(X_1, X_2, \dots, X_N)$, $(Y_1, Y_2, \dots, Y_N)$ sets of real valued random variables; with $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ and $\mathbf{Y} = (\mathbf{y}_1, \dots \mathbf{y}_N)$ their realization. In particular, we consider here a multivariate $\mathbf{X}$, for the principal direction of the poses of a group of joints, such that a random sample of observations $\mathbf{x}_i \in \mathbb{R}^7$. We consider also a multivariate $\mathbf{Y}$ for the PHOG features, with $\mathbf{y}_i \in \mathbb{R}^m$, $m \in \{16, 24, 32\}$. To simplify reading we sometimes talk about poses, though in fact we consider the twists obtained by the principal direction of the set of rototranslations of the joints of a group, with respect to the same joints in the Vitruvian pose, as explained in Sect. 3.

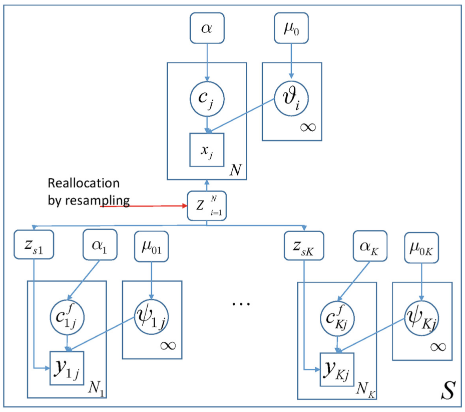

Fig. 5. Plate representation of $S = 1, \dots, 11$ fold replication of the stacked DPM for pose and visual features. Inner plates are replicated for each DPM.

Given the training sets $D^X_s = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^7, i = 1, \dots, N\}$ and $D^Y_s = \{\mathbf{y}_i | \mathbf{y}_i \in \mathbb{R}^m, i = 1, \dots, N, m \in \{16, 24, 32\}\}$ as the sampled pose and visual features for a group $G_s, s = 1, \dots, 11$ (a subset of joints as specified in Fig. 2), we want to partition these sets, though neither the partition dimensions nor the specific allocations are known. Hence we resort to the Bayesian nonparametric perspective on mixtures with countably infinite number of components. In this perspective we are given a measurable space $\mathbb{X}$, a discrete measure μ on this space, a collection of continuous observations, latent variables $(\theta_1, \dots, \theta_K)$ admitting a distribution, with K a random number $\leq N$, and a probability distribution function $F(\cdot|\theta_i)$, parametrized by the random variables θ_i. This setting leads to the popular Dirichlet process mixture model, where $F(\cdot|\theta_i)$ is the kernel of the mixture, here the normal distribution, and $\mu \sim DP(\alpha, \mu_0)$, is the mixing measure, with concentration parameter α and mean $E\{\mu\} = \mu_0$. This is usually expressed in a hierarchical representation as:

$$\begin{array}{ll} X_i|\theta_i = F(\cdot|\theta_i), & i = 1, \dots, K \\ \theta_1, \dots, \theta_N|\mu \sim_{iid} \mu \text{ and} & \mu \sim DP(\alpha, \mu_0). \end{array} \tag{1}$$

Then $X \sim \displaystyle\int F(X|\theta) d\mu(\theta)$ is a mixture of distributions with countably infinite number of components [37,38]. Since the measure μ is discrete, each pair of latent random variables can take the same value with probability $p > 0$. Where the taken value is precisely that of a mixture component. Hence the observations will be allocated by the latent variables to a random number of components.

Different representations have been given of the DPM since [39] and several methods have been devised to sample the mixture parameters from the

$DP(\alpha, \mu_0)$ (see [40,41]). Recently a number of contributions have explored advanced methods to obtain a parallel implementation [42,43], and to obtain a distribution on the partition of the tangent space to the sphere [44], introducing mixture models for data lying on the sphere, and on Riemaniann manifolds [45]. In this work we did not consider our data as placed on a curved manifold. Despite features data for poses are obtained from the principal direction on SE(3), each twist extended with the velocity norm, as described in Sect. 3 is independent of the others and forms an exchangeable set. As we do not consider any trajectory between the pose feature vectors we may not consider them on a curved manifold, though we are exploring the interesting modeling that a manifold representation could lead to. Several approaches have also considered different forms of hierarchical and nested NPB models. Though here we could not use the hierarchical model of [46], since the pose clusters of the same group, likewise the visual features, do not share any element. Neither could be used across groups, since groups have different ranges of PHOG variates and the number of clusters depends on the number of poses of a specific body part.

Our proposed hierarchical model relies on the hypothesis that for the training datasets there exists an index set $\{Z\}_{i=1}^{N}$, with a bijective mapping h between any two datasets. So, for each PHOG feature vector $\mathbf{y}_i$ there exists a corresponding pose vector $\mathbf{x}_i$ in the training set. This fact does not affects generality nor exchangeability, as we see below, since the index set labels the sampled features not the partitions.

To generate an exchangeable random partition for the mixture of poses, we consider the well known Chinese restaurant process (CRP) [47]. On the other hand, to compute the parameter α we followed the approach of [48], defining the prior of α as coming from the class of mixtures of gamma distributions, with small initial scale and shape parameters. For inference we resort to Gibbs sampling [49,50] with conjugate priors.

Given the distribution on the partition induced by the mixture model, a finite set of parameters $\hat{\theta}_1, \ldots, \hat{\theta}_K$ is obtained, together with a cluster indexing $\mathbf{c} = (c_1, \ldots, c_N)$ for each element in the training set. The prediction of a new pose $\mathbf{x}_{N+1}$ is defined by the posterior predictive distribution:

$$p(\mathbf{x}_{N+1}|\mathbf{X}) = \sum_{c_1,\ldots c_{N+1}} \int p(\mathbf{x}_{N+1}|c_{N+1}, \theta) p(c_{N+1}|\mathbf{c}) p(\mathbf{c}, \theta|\mathbf{X}) d\theta \tag{2}$$

Here:

$$p(\mathbf{c}, \theta|\mathbf{X}) = \frac{1}{H} \prod_{k=1}^{K} \mu_0(\theta_k) \prod_{j=1}^{n} F(x_j|\theta_{c_j}) P(c_j), \tag{3}$$

where H is the marginal likelihood of the mixture of Normals given the computed parameters. And, according to the sampling process induced by the CRP, $p(c_{N+1}|\mathbf{c})$ is:

$$p(c_{N+1} = k|\mathbf{c}) = \begin{cases} \dfrac{n_k}{N-1+\alpha} & k \leq K \\ \dfrac{\alpha}{N-1+\alpha} & \text{otherwise} \end{cases} \tag{4}$$

Here n_k is the size of the set of elements in $\mathbf{c}$ having value k. Since poses are continuous and somehow unpredictable, the case that a new pose asks for the initialization of a new cluster has probability greater than zero. However, once the partition is specified, we make it available to the visual inference, recovering the association between the index set $\{Z\}_{i=1}^N$, and each element in each cluster of the dictionary. Because of the label switching problem we prefer to reallocate the indexes $\{Z\}_{i=1}^N$ to the clusters. Hence, for each pair $\hat{\theta}_{c_i} = (\eta_{c_i}, \Sigma_{c_i})$ we sample a number of pose vectors $\{\mathbf{u}\}_{|\mathbf{c}_i|}$, proportional to the current ones from $(\eta_{c_i}, \beta D)$, with $\Sigma_{c_i} = UDU^\top$, and β a filtering parameter. Given the sampled set we find, in the training set D_s^X, the pose vectors $\mathbf{x}$ which minimize the square error, w.r.t. some specific threshold, i.e. $\{\mathbf{x} \in \mathbf{X} | ||\mathbf{x} - \mathbf{u}||_2 \leq \epsilon, \epsilon > 0\}$. This fact allows, at the same time, to regularize the clusters around their mean, and to reallocate the observations into the clusters together with the observation index set $\{Z\}_{i=1}^N$. Therefore according to the model, the induced partition, and the reallocation, given elements $s = \{\mathbf{x}_{s1}, \dots, \mathbf{x}_{sk}\} | \hat{\theta}_{c_j}$ we have that $h^{-1}(s) = z_{sj}$, a subindex set $z_{sj} \in \{Z\}_{i=1}^N$, such that $h(z_{sj}) = \{\mathbf{y}_{s1}, \dots, \mathbf{y}_{sk}\}$, namely it returns a choice of visual features. The subindex z_{sj} specifies which set of features, having index in $\{Z\}_{i=1}^N$ should be allocated to the cluster generated by parameters θ_{c_j}, due to the bijection between the training data. Repeating this for all parameters $\hat{\theta}_{c_j}$, $j = 1, \dots K$, and for each group, a CRP process is computed for each feature set indexed by z_{sj}. The probability measures generating these new set of DPM, are obviously specific for each PHOG feature set. The structure of the hierarchical model is illustrated in Fig. 5. Each feature set indexed by z_{sj} can specify different views of the same pose, and possibly under different lighting conditions. Further, we expect that similar poses of different people, yet belong to the same cluster, and the PHOGs might capture this, when represented by a mixture distribution. Thus we induce a new partition exploiting the Gamma additive property. For each cluster of poses, generated by each group, there exists a set of models $\mathbb{M}_s = (\mathcal{M}_{PHOG}^1, \dots, \mathcal{M}_{PHOG}^{K_s})$, with K varying according to the group $s, s = 1, \dots, 11$.

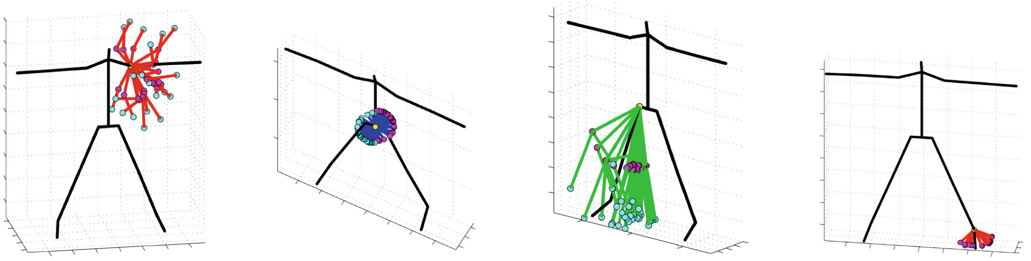

Fig. 6. Most representative poses of the learned dictionary for the groups *Left Arm, Hips, Right Leg, Left Foot*, with respect to the "Vitruvian pose".

Now, given a new observation $\mathbf{y}^\star$, this could be either a query or a new measure. Then the posterior predictive of Eq. (2) should integrate with respect to the parameters of the feature set indexed by z_{sj}, for $j = 1, \dots, K$ and with respect to each feature set $\mathbf{Y}_{z_{sj}}$, collected in the training. Without loss of generality we can do this into two steps. In the first step we compute the density, finding the model that best fits $\mathbf{y}^\star$. We can do this because the index set for the visual features is not required for this step:

$$\arg\max_{\mathcal{M}_{PHOG}} p(\mathbf{y}^\star|\xi) = \sum_h \sum_j \pi_{hj}\varphi_h(\mathbf{y}^\star|\xi_{hj}, \mathcal{M}^h_{PHOG}). \tag{5}$$

Here the πs are the mixing proportion and $\varphi(\cdot|\xi)$ is the Normal density with parameters ξ for the specific PHOG features set. Once the model is chosen, hence the cluster, the predictive distribution in Eq. (2), can be applied to the PHOG feature $\mathbf{y}^\star$. Note that if a new component is generated, this now will have its reference pose being the mean of the cluster it is hooked to. Note that if the subindexes of the clusters generated by the visual features $\mathbf{y}$ with subindex z_{sj} are needed, to identify a particular feature and its connection to a particular pose, then a resampling is necessary, as we did with the poses. Otherwise the mean pose can be used. We can see this process as a funnel guiding visual features into the small opening of the pose set, and possibly widening the opening as new observations come in.

```
Data: Pairwise group compatibility probabilities $r_{ij}$ (Eq. 6).
Result: Most likely set of consistent pose clusters.
Find the most likely pose cluster for the root group ($G_8$);
Add all the connected groups of $G_8$ (denoted $children(G_8)$) in the set $\mathcal{G}_{open}$;
while $\mathcal{G}_{open}$ is not empty do
    for Each group $G_s \in \mathcal{G}_{open}$ do
        Find its most likely pose cluster taking into account the compatibilities $\{r_{ij}\}_{i\in\{1,M_s\}}$ with respect to the selected cluster $j$ of its parent group $parent(G_s)$
    end
    Remove ($G_s$) from $\mathcal{G}_{open}$;
    Add $children(G_s)$ in $\mathcal{G}_{open}$
end
```

Algorithm 2. Consistent pose cluster selection.

The final inference step requires a principle of compatibility amid groups from which derive the consistent pose selection summarized in Algorithm 2. We define the intergroup clusters compatibility as follows. Let i, j, be two clusters from groups q and s. Let $W_{ij} = |z_{qj} \cap z_{si}|$ with $|\cdot|$ the cardinality and let $D_{ij} = z_{qj} \cup z_{si}$ and $p(m_{ij} = 1) = W_{ij}/|D_{ij}|$.

The probability that the two intergroup clusters are compatible is given as:

$$r_{ij} = \frac{p(D_{ij}|m_{ij}=1)p(m_{ij}=1)}{p(D_{ij}|m_{ij}=1)p(m_{ij}=1) + p(D_{ij}|m_{ij}=0)(1-p(m_{ij}=1))} \tag{6}$$

With

$$p(D_{ij}|m_{ij}=1)=\gamma\sum_{D_{ij}}\pi_i\delta_{D_{ij}}(\mathbf{x})+(1-\gamma)\sum_{D_{ij}}\pi_j\delta_{D_{ij}}(\mathbf{x}) \tag{7}$$

Where $\delta_{D_{ij}}(\mathbf{x}) = 1$ if $\mathbf{x} \in D_{ij}$ and zero otherwise, π_i and π_j are the mixing proportions of the DPM of the two clusters, and $0 \leq \gamma \leq 1$ balances the contribution from the two clusters. While, where the two clusters are completely uncorrelated:

$$p(D_{ij}|m_{ij}=0)=\prod_{D_{ij}}\pi_i\pi_j \tag{8}$$

5 Results

Dictionary Learning. As described in Sect. 3, we consider the dataset Human 3.6 M [24] to evaluate our 3D pose estimation algorithm. In order to obtain the dictionaries of the 3D poses we first apply the decomposition of the joints in groups according to Fig. 2 and then compute PGA-based features for each group joints, as described in Sect. 3. As the dataset contains 3D poses synchronized with video frames at a high rate (50 Hz), we subsample with a factor of 12 in order to remove redundant data. Further we compute the PHOG features as described in Sect. 3. The number of clusters generated for each group by the DPM models are reported in Table 2.

Table 2. Number of clusters generated by the DPM models for the PHOG and the PGA-based features for each group of joints.

Groups	1	2	3	4	5	6	7	8	9	10	11	12
Nr. of pose clusters	56	155	38	85	20	49	90	88	58	49	52	16
Avg. nr. of visual components	18	31	31	25	22	22	4	22	22	11	18	13

The significance of pose clusters is shown in Fig. 6, where the mean poses are visualized for the groups *Left Arm, Hips, Right Leg, Left Foot.*

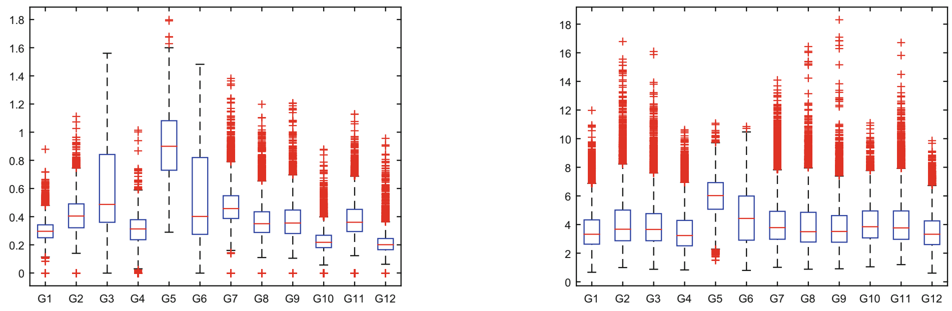

Fig. 7. Error distribution for the PHOG (left) and the PGA (right) features.

3D Pose Estimation. Using the learned dictionary of poses and visual features we perform 3D pose estimation for the testing part of the dataset, namely for the actions performed by subjects S9 and S11. For each query image, the 2D joint positions in the image are estimated by using [2], and they are grouped together forming the groups of Fig. 2. For each group, the PHOG features are then extracted, as described in Sect. 3, and the corresponding cluster of the visual dictionary is selected as the most likely one according to the learned hierarchical model. We calculate the error of the visual features as the euclidean distance of the extracted features with respect to the most representative visual features of the selected cluster. The mean of this error together with the 25th and 75th percentiles for each group, are shown in the left box-plot of Fig. 7. Note that as the errors refer to distances, we expect that they follow a χ^2 distribution instead of a normal one. We observe that the errors of the PHOG features are low in average for most of the groups. The groups corresponding to the hands and the arms (G_3, G_4, G_5, G_6) show higher errors, mainly because of the high variability of their appearance.

The 3D pose of the whole body is obtained according to Algorithm 2.

Table 3. Average per joint error between the estimated 3D pose and the ground truth in mm. Best values in bold.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
LinKDE [24]	132.71	183.55	132.37	164.39	162.12	205.94	150.61	171.31
Li et al. [13]	-	136.88	96.94	124.74	-	168.68	-	-
Tekin et al. [51]	102.39	158.52	87.95	126.83	118.37	185.02	114.69	107.61
Zhou et al. [17]	87.36	109.31	**87.05**	103.16	116.18	143.32	106.88	**99.78**
Ours	**48.82**	**56.31**	95.98	**84.78**	**96.47**	**105.58**	**66.30**	107.41
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	**Average**
LinKDE [24]	151.57	243.03	162.14	170.69	177.13	96.60	127.88	162.14
Li et al. [13]	-	-	-	-	132.17	69.97	-	-
Tekin et al. [51]	136.15	205.65	118.21	146.66	128.11	**65.86**	**77.21**	125.28
Zhou et al. [17]	124.52	199.23	107.42	118.09	**114.23**	79.39	97.70	113.01
Ours	**116.89**	**129.63**	**97.84**	**65.94**	130.46	92.58	102.21	**93.15**

Examples of the recovered poses for query images of the subjects S9 and S11 are shown in Fig. 8. We calculate the euclidean distance of the PGA-based features of the true 3D pose of the subject, with respect to the most representative PGA-based features of the selected cluster for each group. The mean distance for each group together with the 25th and 75th percentiles, are shown in the right box-plot of Fig. 7. We note that the average errors of the PGA-based features are small for all groups, apart from G_5 and G_6 which correspond to the right arm and the right hand. The fact that the PGA features reside in a deeper level of the hierarchical model affects the presence of an increased number of errors above the 95th percentile.

We also compute the mean error of the joint positions of the recovered 3D pose with respect to the ground truth 3D pose of the subject. This error, compared to the error of other state of the art approaches is reported in Table 3.

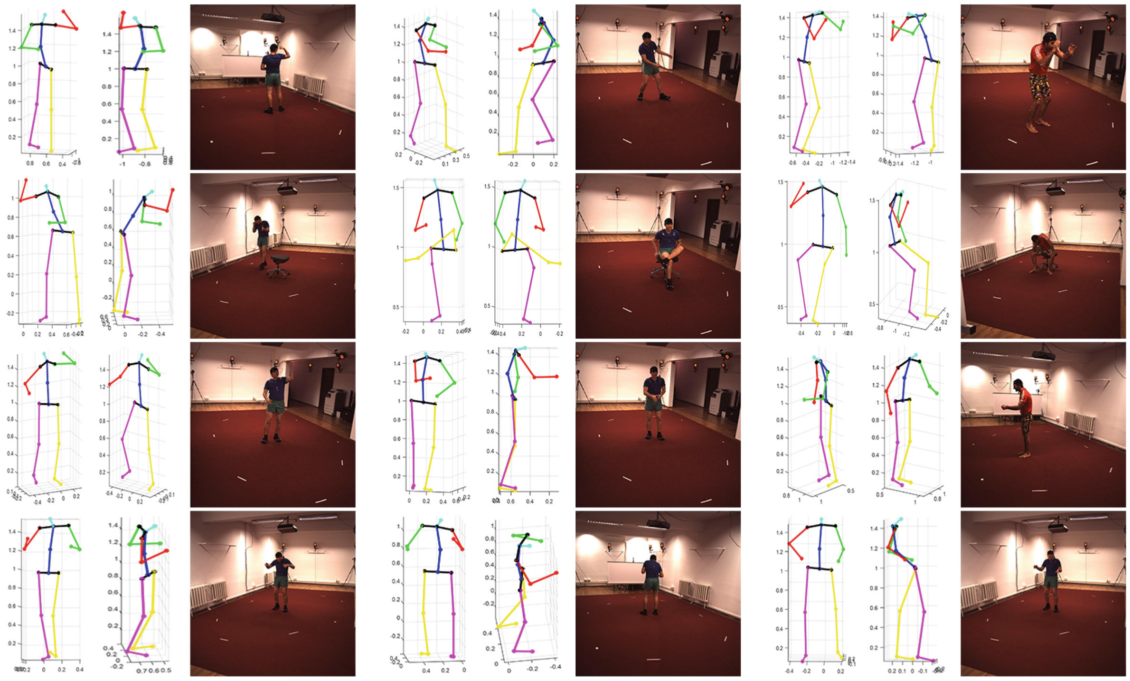

Fig. 8. Examples of query images and the recovered 3D pose. More results are reported in the supplementary material.

The results show that our method gives slightly worse results only with respect to [17] for the 'Eating' and 'Purchases' actions, and for the walking actions with respect to [51] and [17]. In summary, the proposed method outperforms other recently proposed state of the art 3D pose estimation methods both in average and also for the vast majority of actions considered in the Human 3.6 M dataset.

Efficiency of the Method. For the 2D joints estimation training uses 61750 frames of the Human 3.6 M dataset taking about 10^4 s, [2] does not report efficiency. For the hierarchical DPM we consider a training set of 130272 frames, asking for $\sim 8.5 \times 10^5$ s for the poses partitioning and $\sim 7 \times 10^4$ s for the visual features partitioning. This considering main Gibbs cycles of 1800 iterations. Full-pose consistency takes around 0.05 s for a single query, and the total percentage of queries not satisfying it are around 23 %. Once parameters are learned pose computation takes around 0.96 s, with PGA and group computation taking around 0.07 s. These results are obtained with a computer equipped with four Xeon E5-2643, 3.70 GHz CPUs and 64 GB RAM.

6 Conclusions

We present a novel method for 3D human pose estimation from a single image based on a hierarchical Bayesian non-parametric model. The proposed model captures idiosyncratic variations of the motion and the appearance of different body parts, identified by groups of joints. The decomposition in groups avoids redundant configurations, obtaining a more concise dictionary of poses and visual appearances. Given the learned model a 3D pose query can be resolved in real-time. The results show that the proposed model is able to generalize and accurately reconstruct the 3D pose of previously unseen subjects. Our results improve the current state of the art though we aim to further ameliorate them, by considering additional constraints of the pose structure. We shall also consider to move the NBP on the Riemann manifold for the pose features considered.

Acknowledgement. Supported by EU FP7 TRADR (609763) and EU H2020 SecondHands (643950) projects. The authors thank the anonymous reviewers for their comments.

References

1. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the International Conference on Image and Video Retrieval, pp. 401–408. ACM (2007)
2. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2878–2890 (2013)
3. Liu, Z., Zhu, J., Bu, J., Chen, C.: A survey of human pose estimation: the body parts parsing based methods. J. Vis. Commun. Image Represent. **32**, 10–19 (2015)
4. Hen, Y.W., Paramesran, R.: Single camera 3D human pose estimation: a review of current techniques. In: Proceedings of the International Conference for Technical Postgraduates (TECHPOS), pp. 1–8. IEEE (2009)
5. Poppe, R.: Vision-based human motion analysis: an overview. Comput. Vis. Image Underst. **108**(1), 4–18 (2007)
6. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1446–1455 (2015)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005)
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
9. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)
10. Zhou, F., Torre, F.: Spatio-temporal matching for human detection in video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 62–77. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_5
11. Simo-Serra, E., Ramisa, A., Alenyà, G., Torras, C., Moreno-Noguer, F.: Single image 3D human pose estimation from noisy observations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2673–2680. IEEE (2012)
12. Wang, C., Wang, Y., Lin, Z., Yuille, A., Gao, W.: Robust estimation of 3D human poses from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2361–2368 (2014)
13. Li, S., Chan, A.B.: 3D human pose estimation from monocular images with deep convolutional neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9004, pp. 332–347. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16808-1_23

14. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems, pp. 1799–1807 (2014)
15. Ouyang, W., Chu, X., Wang, X.: Multi-source deep learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2329–2336 (2014)
16. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
17. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K., Daniilidis, K.: Sparseness meets deepness: 3D human pose estimation from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2016)
18. Sigal, L., Black, M.J.: Predicting 3D People from 2D Pictures. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2006. LNCS, vol. 4069, pp. 185–195. Springer, Heidelberg (2006). doi:10.1007/11789239_19
19. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Trans. Pattern Anal. Mach. Intell. **28**(1), 44–58 (2006)
20. Mori, G., Malik, J.: Recovering 3D human body configurations using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7), 1052–1062 (2006)
21. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 677–684. IEEE (2000)
22. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 623–630. IEEE (2010)
23. Lehrmann, A.M., Gehler, P.V., Nowozin, S.: A non-parametric Bayesian network prior of human pose. In: Proceedings of the IEEE International Conference on Computer Vision (2013)
24. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human 3.6m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
25. Pons-Moll, G., Fleet, D., Rosenhahn, B.: Posebits for monocular human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2337–2344 (2014)
26. Yasin, H., Iqbal, U., Krüger, B., Weber, A., Gall, J.: 3D pose estimation from a single monocular image. arXiv preprint arXiv:1509.06720 (2015)
27. Flaherty, F., do Carmo, M.: Riemannian Geometry. Mathematics: Theory and Applications. Birkhäuser, Boston (2013)
28. Zefran, M., Kumar, V., Croke, C.: On the generation of smooth three-dimensional rigid body motions. IEEE Trans. Robot. Autom. **14**, 576–589 (1998)
29. Duan, X., Sun, H., Peng, L.: Riemannian means on special euclidean group and unipotent matrices group. Sci. World J. **2013** (2013). doi:10.1155/2013/292787
30. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The vitruvian manifold: inferring dense correspondences for one-shot human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 103–110. IEEE (2012)
31. Fletcher, P., Lu, C., Pizer, S., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imaging **23**, 995–1005 (2004)
32. Karcher, H.: Riemannian center of mass and mollifier smoothing. Commun. Pure Appl. Math. **30**(5), 509–541 (1977)

33. Afsari, B., Tron, R., Vidal, R.: On the convergence of gradient descent for finding the riemannian center of mass. SIAM J. Control Optim. **51**(3), 2230–2260 (2013)
34. Kendall, W.S.: Probability, convexity, and harmonic maps with small image. I: uniqueness and fine existence. Proc. Lond. Math. Soc. **3**(2), 371–406 (1990)
35. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. Int. J. Comput. Vis. **103**(3), 267–305 (2013)
36. Natola, F., Ntouskos, V., Pirri, F., Sanzari, M.: Bayesian non-parametric inference for manifold based MoCap representation. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 4606–4614 (2015)
37. Lo, A.Y.: On a class of Bayesian nonparametric estimates. I: density estimate. Ann. Statist. **12**, 351–357 (1984)
38. Ferguson, T.: Bayesian density estimation by mixtures of normal distributions. Recent Adv. Stat. **1**, 287–302 (1983)
39. Ferguson, T.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)
40. Gorür, D.: Nonparametric Bayesian discrete latent variable models for unsupervised learning. Ph.D. thesis, Max Planck Institute for Biological Cybernetics (2007)
41. Sudderth, E.B.: Graphical models for visual object recognition and tracking. Ph.D. thesis, MIT (2006)
42. Lovell, D., Adams, R.P., Mansingka, V.: Parallel markov chain monte carlo for Dirichlet process mixtures. In: Workshop on Big Learning, NIPS (2012)
43. Chang, J., Fisher III., J.W.: Parallel sampling of DP mixture models using subcluster splits. In: Advances in Neural Information Processing Systems, pp. 620–628 (2013)
44. Straub, J., Chang, J., Freifeld, O., Fisher III., J.W.: A Dirichlet process mixture model for spherical data. In: AISTATS (2015)
45. Kim, H.J., Xu, J., Vemuri, B.C., Singh, V.: Manifold-valued Dirichlet processes. In: Proceedings of the 2015 International Conference on Machine Learning, vol. 2015, pp. 1199–1208 (2015)
46. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**, 1566–1581 (2012)
47. Pitman, J.: Combinatorial Stochastic Processes: Ecole D'Eté de Probabilités de Saint-Flour XXXII-2002. Springer, Heidelberg (2006)
48. West, M.: Hyperparameter estimation in Dirichlet process mixture models. ISDS discussion paper# 92–A03, Duke University (1992)
49. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)
50. Jain, S., Neal, R.M.: A split-merge Markov chain monte carlo procedure for the Dirichlet process mixture model. J. Comput. Graph. Stat. **13**, 158–182 (2012)
51. Tekin, B., Sun, X., Wang, X., Lepetit, V., Fua, P.: Predicting people's 3D poses from short sequences. arXiv preprint arXiv:1504.08200 (2015)