Skeletal pose-priors for improved human body fitting to scans

ROURE Bastien

Supervised by: Sergi PUJADES[1], Abdelmouttaleb DAKRI[1]
[1]Inria, Montbonnot, France

Abstract

The aim of this work is to provide a way of distinguishing the plausibility of 3D human body model poses using what is called **pose-prior** : statistical models that evaluate the likelihood of a pose being realistic. Various pose priors exist for models that originate from the computer graphics community, such as SMPL, but are non-existent for a class of more accurate and recent bio-mechanical models such as **SKEL**. To address this lack, we use a **Gaussian Mixture Model** (GMM) to introduce the first **SKEL pose-prior**. To provide a data-set for the GMM a set of realistic human SKEL poses is needed, a **SKEL data-set** was therefore created. We then extensively evaluate the GMM on various tasks. The resulting prior enables more accurate results for SKEL registrations to scans.

1 Introduction

The modeling of the shape and motion of human bodies is a field of research with a wide range of applications, from films, animations and video games, to bio-mechanics and medicine. One of the main challenges within this field is to represent different body shapes and ensure that the movements associated with them are realistic (constraints on joints, anatomically accurate representations of the inner tissues and the skeleton, soft-tissue deformations, ...).

In order to capture humans in motion, one of the most widely used data acquisition techniques is motion capture (MoCap), as it offers a good compromise between efficiency, realism of movement, and cost. It involves recording the positions of markers, placed on the body, by cameras or sensors. These recorded points form a time-varying spatial data-set, which is then structured into a kinematic tree. The MoCap thus provides rich information on movement dynamics, albeit being relatively poor in terms of appearance details or surface geometry. Alternatively, another way of data acquisition is 3D scanning. Techniques like multi-view stereo, photogrammetry, enable detailed shape capture at a higher cost. Unlike MoCap however, it faithfully captures information on the shape surface and texture.

Once these data have been collected, a compact, controllable, and differentiable 3D representation of the human body is required, enabling applications like pose estimation, animation, and simulation. Parametric body models are a popular solution that addresses these tasks, using a small number of parameters to generate both realistic shapes and poses. Body models also function as human prior on 3D data: since captured data is generally unstructured and semantically ambiguous, parametric body models play the role of a proxy to the 3D input data. Without a body model, a mesh is just an unordered collection of triangles and vertices!

To be able to adapt or register a parametric model to data retrieved by MoCap or scans, one has to be able to register said model into the input geometry. These registrations use proximity-based optimizations to minimize the distance between the model and the data, searching through a large space of shape and pose parameters. So in order to obtain plausible results through the optimization, and therefore to accurately represent the input data, it is necessary to be able to distinguish the poses that a human can have from those that are unrealistic and penalize the latter within the optimization loop.

Statistical **pose-priors**, which attempts to model the likelihood of a given pose being plausible, are extensively used in the literature for graphics models such as SMPL (Skinned Multi-Person Linear) [Loper *et al.*, 2015], the likes of VPoser [Pavlakos *et al.*, 2019], DPoser [Lu *et al.*, 2023], Pose-NDF [Tiwari *et al.*, 2022] and NRDF [He *et al.*, 2024]. They can be seen as a function which, taking a pose, returns the probability that it is realistic. However, for more recent biomechanically inspired models, such as **SKEL** [Keller *et al.*, 2023], no prior exists.

The aim of this internship is therefore to provide a **pose-prior** for the bio-mechanical human body model **SKEL** and evaluate its effectiveness on scan registration tasks. To do so, we propose a new large data-set of **SKEL** poses. As such, the main contributions of this paper are :

- A large data-set of SKEL poses adapted from pre-registered scan and MoCap data.
- A statistical prior for **SKEL** poses using a **Gaussian Mixture Model** trained on the aforementioned data-set.
- An evaluation of the pose-prior on tasks like random pose disambiguation and scan registration.

2 Related Work

2.1 Human Body Models

There are many parametric models, such as SCAPE [Anguelov *et al.*, 2005] and GHUM [Xu *et al.*, 2020]. One of the most widely used models within the computer vision community is **SMPL** and its derivatives: STAR [Osman *et al.*, 2020], SMPL-X [Pavlakos *et al.*, 2019] and SUPR [Osman *et al.*, 2022].

SMPL uses an artist-created template mesh with a fixed topology $N = 6890$ vertices, and $K = 23$ joints. It represents any specific body shape and pose as a deformation of this template. The model is defined by a vector of N vertices $\bar{\mathrm{T}} \in \mathbb{R}^{3\times N}$, in a standard pose θ^* (Figure 1(a.)/(b.)/(c.)); a set of blend weights $\mathcal{W} \in \mathbb{R}^{N\times K}$ Figure 1(a), that represent the influence of joints on each point; a blend shape function, $B_S(\beta) : \mathbb{R}^{|\beta|} \mapsto \mathbb{R}^{3N}$, which, taking β , the shape parameter, returns the deformation of the body shape Figure 1(b). $J(\beta) : \mathbb{R}^{|\beta|} \mapsto \mathbb{R}^{3\times K}$ is a function which, with the same β parameter, predicts the location of the K joints as a function of body shape (white dots in Figure 1(b)). $B_P(\theta) : \mathbb{R}^{|\theta|} \mapsto \mathbb{R}^{3\times N}$, a pose-dependent blend shape function, that takes as input θ a vector of pose parameters, adds deformations depending on the pose (e.g. skin folds) Figure 1(c). A standard blend skinning function $W(\cdot)$ is applied to rotate the vertices around the estimated joint centers with weighting defined by the blend weights.

To summarize :

$$\mathrm{SMPL}(\beta,\theta) = W(T_P(\beta,\theta), J(\beta), \theta, \mathcal{W})$$
$$T_P(\beta,\theta) = \bar{T} + B_S(\beta) + B_P(\theta)$$

Although SMPL (and its derivatives) is a standard for 3D modeling of the human body, it does have its limitations. The model is not suited to accurately represent the physical behavior of the human body. The joints in SMPL are misplaced in relation to those of a human skeleton, but also, modeled with three degrees of freedom, meaning they don't always reflect real anatomical constraints. Models based on the same framework as SMPL are available for obtaining accurate biomechanical skeletal movements.

2.2 Biomechanical Human Body models

The **SKEL** model [Keller *et al.*, 2023], specifically designed to approach better internal behavior (under the skin) by introducing a skeleton BSM (Bio-mechanical Skeleton Model). This model aims to overcome some of the limitations of SMPL, particularly regarding the kinematic realism and anatomical plausibility of joint positions and movements.

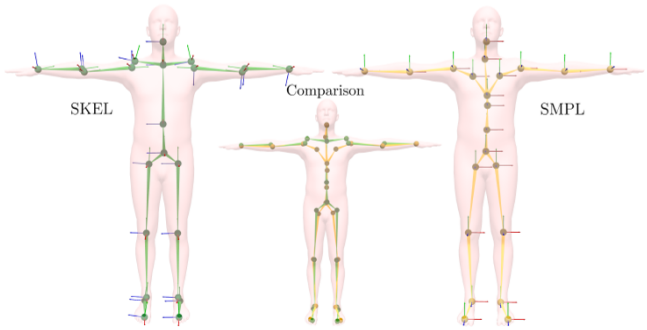

Figure 2: [Image from SKEL paper] Difference between joint placement and resulting kinematic tree between SKEL and SMPL

The SKEL$(\boldsymbol{\beta}, \boldsymbol{q})$ model inherits the same space $\boldsymbol{\beta}$ from SMPL, uses $\boldsymbol{q} \in \mathbb{R}^{46}$, the pose vector from the anatomically accurate kinematic tree (Figure 2) of the skeletal model that consists of 24 rigid groups of bones with joints defined between them as well as a mesh representation of these bones. The main changes introduced in the model definition are : the joints location depending on the shape vector $J^{\mathrm{SKEL}}(\beta) : \mathbb{R}^{|\beta|} \mapsto \mathbb{R}^{3K}$, the new blend weights $\mathcal{W}^{\mathrm{SKEL}} \in \mathbb{R}^{46}$ associated for each point and $B_P(q) : \mathbb{R}^{|q|} \mapsto \mathbb{R}^{3N}$ the pose-dependent blend shape function.

SKEL links the SMPL mesh and a skeleton given by the BSM model, in a single repository, to obtain a more accurate human biomechanical model.

To summarize :

$$\mathrm{SKEL}(\beta,q) = W(T_P(\beta,q), J^{\mathrm{SKEL}}(\beta), q, \mathcal{W}^{\mathrm{SKEL}})$$
$$T_P(\beta,q) = \bar{T} + B_S(\beta) + B_P(q)$$

2.3 Statistical Pose Prior

In statistics, the *prior probability distribution* of an uncertain quantity is the probability distribution assumed before certain evidence is taken into account. In the context of this internship, the **pose-prior** is based on what we consider to be plausible human poses, within the constraints imposed by the biomechanics and anatomy of the human body.

This **pose-prior** is used to optimize the **SKEL** parameters to fit a scan. The general idea is to use a scan of a human body as a starting point, to approximate a mesh defined by parameters from **SKEL** as closely as possible. In this optimization, it is important to ensure that the result is a model that respects the constraints imposed by the human body, as this notion of "proximity" may not be appropriate for the realism of a pose. The **pose-prior** is therefore used as a regularization term to ensure the realism of human poses. We refer the reader to section 6.4 for details about the registration process.

3 Data-set

Accurate modeling of human body movements is a complicated task, which in this case is based on data acquired from

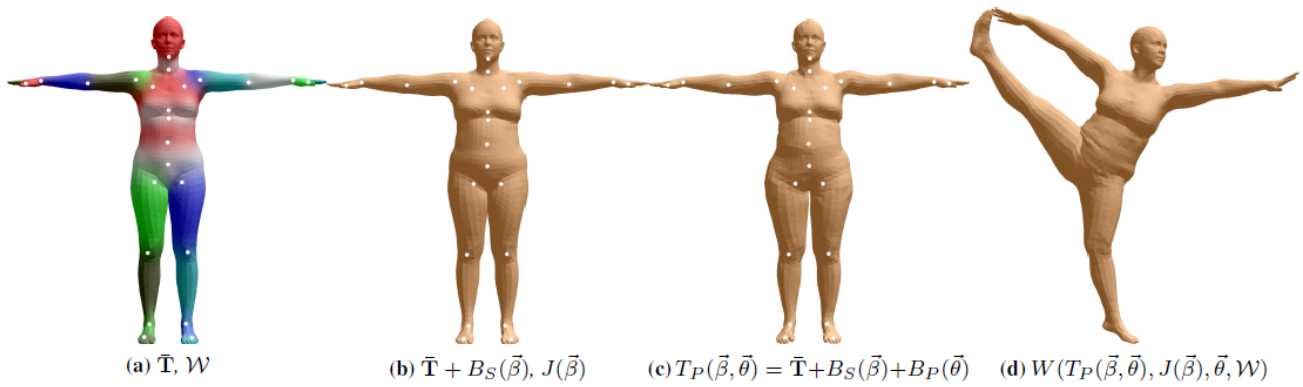

Figure 1: [Image from SMPL paper] **SMPL model** : (a) Template mesh with blend weights indicated by color and joints shown in white. (b) With identity-driven blend shape contribution only; vertex and joint locations are linear in shape vector β. (c) With the addition of pose blend shapes in preparation for the split pose; note the expansion of the hips. (d) Deformed vertices.

real subjects. This study begins with the AMASS (Archive of Motion Capture As Surface Shapes)[Mahmood *et al.*, 2019] and Dynamic FAUST (Fine Alignment Using Scan Texture) [Bogo *et al.*, 2017] data-set, two standardized data-sets of body capture. AMASS is based on MoCap sequences that were converted through regression into SMPL parameters (and therefore SMPL body meshes). FAUST, on the other hand, is based on human body scans that were registered with the help of fiducial markers on the body (a dense coverage of the body with marker stamps). FAUST additionally provides semantically meaningful sequences as motion annotations are available. By combining the large-scale motion diversity of AMASS with the smaller but varied pose sequences of Dynamic FAUST, the result is a comprehensive data-set that captures a wide spectrum of realistic human poses and movements (Figure 3). The TSNE plot (Figure 4) shows that **the data-set is very diverse**, which is essential for the generalization of realistic human poses. A **SKEL data-set** was generated from these data, since SKEL is based on a common mesh topology with SMPL. This structural compatibility is used to make point-to-point matches in a fitting optimization with a loss defined by :

$$\beta^*, q^* = \text{argmin}_{\beta,q}\, d(V_{\text{SKEL}}, V_{\text{SMPL}}) = \text{argmin}_{\beta,q} \sum_{i=1}^{N} \|x_i - y_i\|_2$$

Where $x \in V_{\text{SKEL}}$, are the vertices of the SKEL(β, q) mesh and $y \in V_{\text{SMPL}}$ the vertices of the input SMPL(β, θ) mesh. This loss is optimized using gradient descent, with the *Adam* optimizer [Kingma and Ba, 2017], to find the parameters β and q which minimize this Loss. It's important to note that it's easy to fit a SKEL to a SMPL mesh due to the fact that both meshes share the same topology. This is not the case when this fitting is done to a scan, as the topology of the scan mesh is different and there is no predefined match at every point (as scans are unstructured).

This optimization procedure enables us to obtain a data-set of 400k SKEL poses (see Figure 4), composed of around 130 sequences of an average of 250 frames for DFAUST and 330 sequences of 1220 frames for AMASS, while preserving the consistency of the original sequences. This data-set forms the basis for the **pose-prior** training on SKEL.

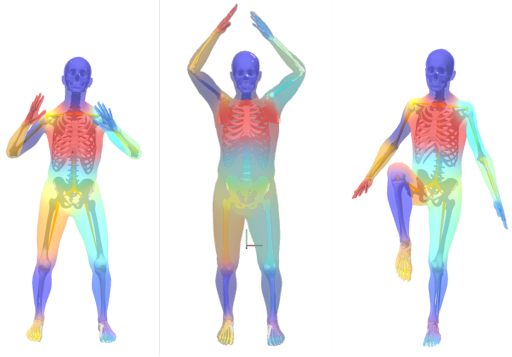

Figure 3: Excerpt of the poses from the SKEL data-set.

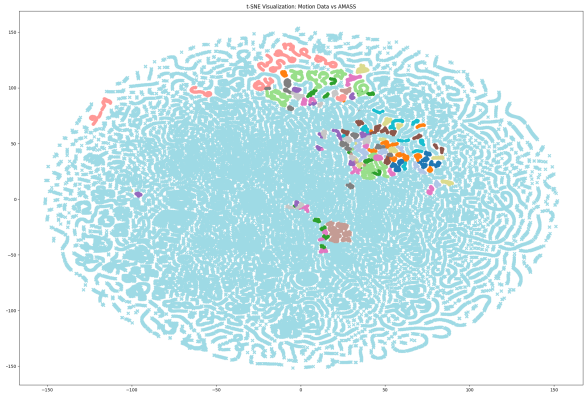

Figure 4: The t-distributed Stochastic Neighbor Embedding plot of our data-set into a 2D space, shows the intrinsic structure of the data-set in terms of postural similarity (AMASS in light blue, DFAUST in other colors, one for each different motion) : two points close together in the plane correspond to similar poses (in terms of cosine similarity of pose vectors).

4 Working Basis

4.1 Gaussian Mixture Model

GMM is a probabilistic model that assumes all data points are generated from a mixture of several Gaussian distributions with unknown parameters. Unlike hard clustering methods like *K-Means*, **GMMs** performs soft clustering, meaning each data point belongs to multiple clusters with certain probabilities defined as follows :

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

with :

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = c \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

and :

$$c = \frac{1}{(2\pi)^{d/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}}$$

where in this case :

- $x \in \mathbb{R}^{46}$: Parameter of a body pose
- π_k : The weight of the k-th Gaussian (s.t $\sum_{k=1}^{K} \pi_k = 1$)
- $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: The Gaussian distribution with mean $\mu_k \in \mathbb{R}^{46}$ and covariance $\Sigma_k \in \mathbb{R}^{46 \times 46}$

Expectation-Maximization (EM) Algorithm is used to fit a **GMM** to the data. This iterative method optimizes the parameters of the Gaussian distributions like mean (μ), covariance (Σ) and weight coefficients (π). Here's how the **GMM** procedure works :

1. **Initialization :** Start with initial guesses for the means, covariances and weight coefficients of each Gaussian distribution.
2. **Expectation step :** The probability $\gamma_{i,k}$ of a point x_i belonging to the cluster k is calculated with the actual parameters.
$$\gamma_{i,k} = P(z_i = k \mid x_i) = \frac{\pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}$$
3. **Maximization step :** Using the probability calculated in the previous step, the algorithm updates the parameters of the model to better fit the data.
$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_{i,k} \qquad \mu_k = \frac{\sum_{i=1}^{N} \gamma_{i,k} \cdot x_i}{\sum_{i=1}^{N} \gamma_{i,k}}$$
$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma_{i,k} \cdot (x_i - \mu_k)^2}{\sum_{i=1}^{N} \gamma_{i,k}}$$
4. **Repeat :** Continue alternating between the **E-step** and **M-step** until the model parameters change insignificantly or equivalently the log-likelihood of the data (which measures how well the model fits the data) converges.

$$\log \mathcal{L} = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right)$$

By training the **GMM** on **SKEL** pose data, it will then give a statistical representation of realistic human body pose configurations. A metric, named ***Mahalanobis Distance***, allows measuring the distance between a point and a Gaussian distribution. This metric is derived to provide a score measuring how closely the point is to the distribution defined by the GMM. It is given by :

$$D_m(x) = \min_{i \in N} \left(\sqrt{(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)} \right)$$

Here the result is the distance between $x \in \mathbb{R}^{46}$, a pose vector, and the nearest Gaussian contained in the **GMM**.

5 Training

We have noticed that at the start of each motion of the data-set, the subjects started in the same position (a standard T pose). To avoid overfitting in this pose, a pre-processing of the data-set was made by keeping the central 80% of each motion sequence in order to concentrate on the most informative part of the motions. With this new data-set made up of the remaining 80%, a 70/30 split train/test was made. This represents training on around 250k poses.

6 Evaluation.

In many clustering algorithms, the choice of the number of clusters (and therefore, in this case, the Gaussians) is a crucial step. This choice has a direct influence on the quality of the data representation, and therefore on the ability to generalize the distribution of the data. There are methods for estimating the right number of clusters, which is a difficult hyper-parameter to estimate for high-dimensional data, that cannot be easily visualized. For a **Gaussian Mixture Model**, one of these methods is the **Bayesian Information Criterion** (BIC). A widely used method for estimating the optimal number of Gaussians. It is defined as follows :

$$\text{BIC} = -2 \cdot \log(\hat{L}) + \log(N)d$$

Where :

$$\log(\hat{L}) = \frac{1}{N} \cdot \log \left(\prod_{i=1}^{N} \cdot \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right)$$

The average likelihood of the data to the GMM distribution. This value increases with the number of components in the model. Indeed, the more Gaussian components the model has, the greater its ability to faithfully approximate the data distribution. Note : the -2 factor is a statistical convention related to the theory of likelihood tests.

N represents the number of elements and d is the degree of freedom of the GMM.

Where :

- **Weights** : As a reminder defined by $\pi_1, \pi_2, ..., \pi_k$ with the constraint

$$\sum_{k=1}^{K} \pi_k = 1$$

So that K-1 weights are free, the last one is determined by the constraint

- **Means** : $\mu_k \in \mathbb{R}^m$ (in this case $m = 46$), making a total parameter of $K \times m$

- **Covariances** : $\Sigma_k \in \mathbb{R}^{m \times m}$ symmetric matrices, meaning that it contains $\frac{m(m+1)}{2}$ independent parameters.

The final result is :

$$d = \underbrace{(K-1)}_{\text{weights}} + \underbrace{K \times m}_{\text{means}} + \underbrace{K \times \frac{m(m+1)}{2}}_{\text{covariances}}.$$

The purpose of the second term $\mathbf{log}(\boldsymbol{N})\boldsymbol{d}$ is to penalize the model if it becomes too complex, in order to avoid overfitting. This criterion is therefore a compromise between the quality of the fit to the data and the simplicity of the model.

6.1 Synthetic Example

In order to validate our implementation, we proceed with a toy test in which we test our training and evaluation pipeline on a known Gaussian Mixture distributed toy sample.

To better illustrate the behavior of BIC as a function of the number of Gaussian components, here's a simple example.

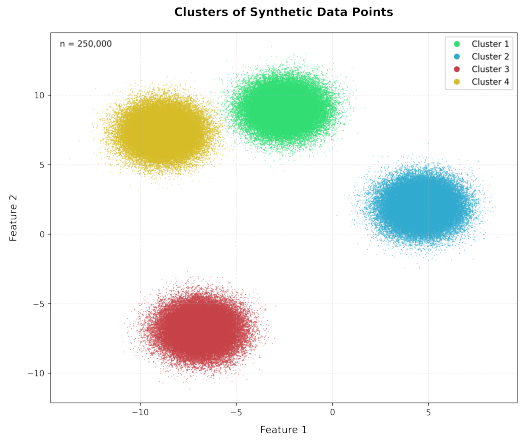

Figure 5: Synthetic data created by 4 Gaussian point distributions, of the same order of magnitude as the SKEL data-set (around 250k) in 2D for visualization.

In this example (Figure 5), it's easy to distinguish 4 clusters. To ensure that the indicator is working properly, it should indicate this number.

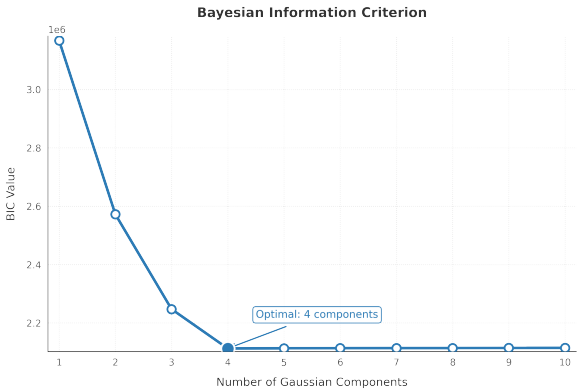

Figure 6: BIC analysis on synthetic data-set.

These results on synthetic data (Figure 6) confirm that BIC can identify a relevant number of components when the data do indeed follow a Gaussian structure.

6.2 SKEL data.

The same test is carried out with data from the **SKEL** data-set, whose underlying structure is a priori more complex.

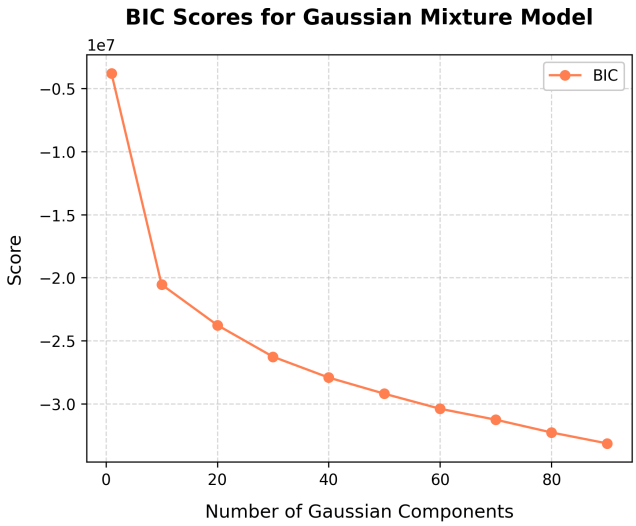

Figure 7: BIC analysis on SKEL data-set.

In this result (Figure 7), the indicator shows no relevant Gaussian number up to 90. Tests were not carried out on a larger number, as the computational costs are prohibitive. In comparison with a SMPL GMM baseline which uses only 6 gaussians, the optimal number of gaussians seems larger. Note however, that since the SMPL baseline has been trained on different data-set (the details of which weren't released to the public) we couldn't carry out a similar analysis on it. Our working hypothesis is that our data-set does not follow the distribution of a mixture of Gaussians and therefore that the model's capacity to represent the data is very limited, making prone to overfit on the data with an increasing number of gaussians. However, in optimization what is needed is a robust method to distinguish between unrealistic and realistic

poses ie to provide a robust score function that penalizes unrealistic SKEL poses. We proceed to show that a GMM trained on the SKEL data-set with 13 gaussians (chosen heuristically) can fulfill this task.

6.3 Other Approach.

Since the aim was for the Gaussian Mixture Model to represent the distribution of realistic poses relatively to unrealistic poses, another approach was to define a data-set **RAND** composed of vectors $x^R \in \mathbb{R}^{46}$ random poses (unrealistic) and compare the score given by the derived Mahalanobis distance $\boldsymbol{D_m}$ and the one given on the **TEST** data-set defined by $x^T \in \mathbb{R}^{46}$ the remaining 30% poses from the starting data-set.

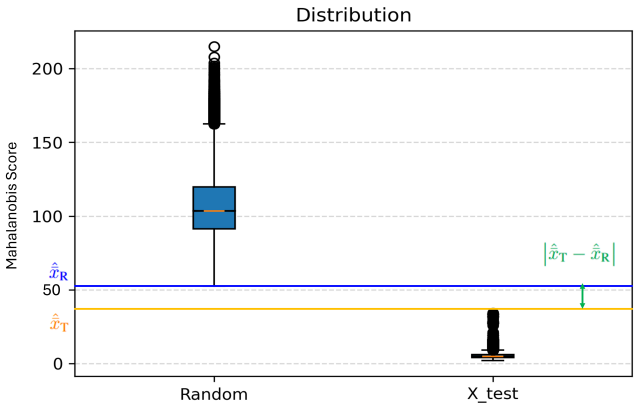

Figure 8: Boxplot illustrating the distribution of the Random data-set and the Test set with 13 Gaussians

With these informations (Figure 8):

- $\hat{\hat{x}}_R = \min_{x_i^R \in \text{RAND}} (D_m(x_i^R))$: The minimum score given on the pose in **RAND**
- $\hat{\hat{x}}_T = \max_{x_i^T \in \text{TEST}} (D_m(x_i^R))$: The maximum score given on the pose in **TEST**

The objective is to have $x_i^R > \hat{\hat{x}}_T, \forall x_i^R \in$ RAND, to avoid false-positive and maximise $|\hat{\hat{x}}_T - \hat{\hat{x}}_R|$

The choice of 13 Gaussians was made on the basis of the result of a maximisation test of A on a number of Gaussians ranging from 1 to 15, with computational cost in mind.

6.4 Results

We evaluate our model on the applied task of scan registration. Given an input scan $\mathcal{S}$, and approximate initial poses and shapes β_0, q_0 (which can be initialized randomly, based on usual predefined shapes and poses or using advanced techniques like image based pose estimation) we aim to solve the following non linear optimisation problem

$$\beta^*,\ q^* = \text{argmin}\ \mathcal{L}(\mathcal{S}, \text{SKEL}(\beta, q); \beta, q) + \lambda \mathcal{L}_{\text{reg}}(\beta, q)$$

where $\mathcal{L}(\mathcal{S}, \text{SKEL}(\beta, q); \beta, q) = \|\mathcal{S} - \text{SKEL}(\beta, q))\| = \sum_{x \in \text{SKEL}(\beta,q)} \text{argmin}_{y \in \mathcal{S}} \|y - x\|$; is the proximity loss that minimizes the distance between the scan and the model at a given optimization step. The distance is based on nearest neighbor euclidean distance between the scan and the model vertices. $\mathcal{L}_{\text{reg}}$ is a set of regularization losses that penalizes deviations from the model learned parameters. For example, as the shape parameters of the SKEL model are based on normalized PCA, a deviation beyond $[-1, 1]$ indicates the model shape at the given optimization step is out of the distribution of the SKEL/SMPL shape space. In this case, an L2 regularization on the shape parameter $\mathcal{L}_{\text{reg}}^{\beta} = ||\beta||_2$ constrains it's variation beyond those limits. We choose to focus on the GMM prior regularization loss. We define our loss as the minimum log likelihood of a given pose with regards to the learned GMM:

$$\mathcal{L}_{\text{reg}}^{q}(q) = min_k\ w_k \times (q - \mu_k)^T \Sigma_k^{-1} (q - \mu_k)$$

Where k is the k^{th} Gaussian, μ_k, Σ_k w_k the associated mean, covariance matrix and weight. We show that the use of a pose prior is essential in obtaining realistic poses as a result of the registration (Figure 9).

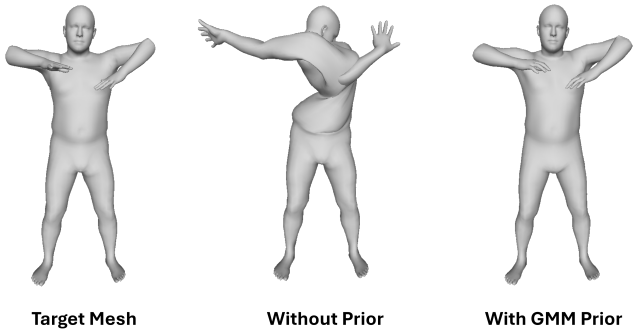

Figure 9: Impact of the SKEL prior on the optimization result.

7 Conclusion

In this work, we proposed the first pose prior for the SKEL model based on a Gaussian Mixture Model, to approximate the distribution of plausible human poses. By integrating this prior into a registration task, we were able to improve the quality of the registered poses.

However, our approach has certain limitations inherent in the use of GMM's, notably a restricted ability to model complex and highly non-linear distributions. These limitations pave the way for the exploration of more expressive methods, such as DPoser [Lu *et al.*, 2023], NRDF (Neural Riemannian Distance Fields) [He *et al.*, 2024], which are potentially better adapted to the structure of the human pose space. This approach is a first step towards building more powerful and flexible prior for the SKEL model.

References

[Anguelov *et al.*, 2005] Anguelov, Srinivasan, Koller, Thrun, Rodgers, and Davis. Scape: shape completion

and animation of people. *ACM Transactions on Graphics (TOG)*, 2005.

[Bogo *et al.*, 2017] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[He *et al.*, 2024] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[Keller *et al.*, 2023] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3D digital humans. *ACM Transaction on Graphics (ToG)*, 42(6):253:1–253:15, December 2023.

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[Loper *et al.*, 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[Lu *et al.*, 2023] Junzhe Lu, Jing Lin, Hongkun Dou, Yulun Zhang, Yue Deng, and Haoqian Wang. Dposer: Diffusion model as robust 3d human pose prior. *arXiv preprint arXiv:2312.05541*, 2023.

[Mahmood *et al.*, 2019] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[Osman *et al.*, 2020] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020.

[Osman *et al.*, 2022] Ahmed A A Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human body model. In *European Conference on Computer Vision (ECCV)*, 2022.

[Pavlakos *et al.*, 2019] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[Tiwari *et al.*, 2022] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, October 2022.

[Xu *et al.*, 2020] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum ghuml: Generative 3d human shape and articulated pose models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Oral)*, pages 6184–6193, 2020.