3D Shape Sequence of Human Comparison using Current and Varifolds (Supplementary Material)

Emery Pierson[1], Mohamed Daoudi[1,2], and Sylvain Arguillere[3]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
emery.person@univ-lille.fr
[2] IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems
mohamed.daoudi@imt-nord-europe.fr
[3] Univ. Lille, CNRS, UMR 8524 Laboratoire Paul Painlevé, Lille, F-59000, France
sylvain.arguillere@univ-lille.fr

1 Comparison with state-of-the-art

In this section, we explain in more details the state-of-the-art methods. Extensive comparison has been made in [9,7] to evaluate the descriptors for human motion retrieval on CVSSP3D dataset. The polygonal curves of those descriptors are filtered with a temporal filtering approach (a mean filter is applied along a temporal window of size K). Finally, the dynamic time warping distance is used for comparing the resulting curves. We compare our motion retrieval approach to the best features presented in those papers, and to several other learned descriptors:

1. The 3D harmonics descriptor [6][9] is a descriptor based on point cloud repartition in space. A 3D shape is first normalized with two variations of PCA. Then, a spherical histogram with different rays is built. The final descriptor is decomposed along spherical harmonics of the obtained with a specific re-weighting for better results. Temporal filtering is proposed in order to deal with the real dataset. We report the results from [9].
2. Breadths spectrum and shape invariant [7] are presented as 2 fully invariant descriptors derived from convex shape analysis. The authors propose to use the breadths of the projection of a shape along each axis spanned by a normal $u \in \mathbb{S}^2$ and to keep the rotation invariant spherical spectrum as a descriptor for human pose. They combine the proposed descriptor with weighted areas of the projection on each plan spanned by u to build a shape invariant. Noise robust version of this descriptor, along with specific temporal filtering named Q-breadths and Q-shape invariant are proposed for the real dataset.
3. Areas, Breadths are the full spherical signals of breadths and weighted areas is proposed to deal with dataset that shows no rotations in [7]. We apply those descriptors, of size 64, along with their concatenation, Areas & Breadths, on Dyna dataset.
4. Aumentado-Armstrong *et al.* [1] propose a variational autoencoder called Geometrically Disentangled VAE (GDVAE). They use PointNet architecture as

point cloud encoders and decoders. In the paper, the authors propose to use disentangled intrinsic and extrinsic latent vectors for human shape representation. PointNet encoder is parameterization invariant, but training loss uses the mesh Laplace Beltrami operator which needs a constant parameterization along the training set. Constraints are applied in training to make the network rotation invariant. We report the result of their extrinsic latent vectors (belonging to $\mathbb{R}^{12}$) from [7]. The network was pretrained on the SURREAL dataset [8]. For the CVSSP3D datasets, we report the results from [7].

5. Zhou *et al.* [10] propose a mesh autoencoder based on the Neural3DMM [3] mesh neural network architecture. The network is only applied on human shapes, with the objective to disentangle shape and pose in latent space. The network architecture requires that all input meshes have the same parameterization. We can thus apply it only on the artificial dataset. We report the cross validated results from [7] using the pose latent vectors (belonging to $\mathbb{R}^{112}$) in the human sequence retrieval pipeline. Since the input of the network are the coordinates of the vertices, the approach is not rotation invariant. For the artificial dataset, we report the results from [7].
6. Cosmo *et al.* [4] propose a similar approach as GDVAE, called Latent Interpolation with Metric Priors (LIMP). They use the same type of autoencoder as GDVAE but change the disentanglement constraints with metric prior constraints: a change in extrinsic latent space should only induce change on extrinsic distances of the meshes, while a change in intrinsic latent space should only induce change on intrinsic distances of the meshes. They use Euclidean and geodesic pairwise matrices in their losses to model this constraint, which needs a constant parameterization in the training set. We use the network pretrained on the FAUST dataset [2]. They do not make any specific training for Euclidean invariance. In order to do motion retrieval, we applied the meshes as input of their available trained network and gathered their extrinsic latent vectors (belonging to $\mathbb{R}^{64}$), and used them in the human sequence retrieval pipeline.
7. Skinned Multi-Person Linear model (SMPL) pose representation. The SMPL body model [5] is a parameterized human body model. A template is deformed (non-rigidly) according to a deformation parameterized by a shape vector. A skeleton is associated to this template and a pose vector, composed of relative rotation of each skeletal joint compared to its parent joint. We convert each joint rotation to quaternion representation as in [10,1] and measure the distance between unit quaternions by $d(q, q') = 1 - |q.q'|$. The SMPL body pose vector contains the pose information of 20 joints, and the rotation of the central joint accounts for the global rotation of the shape, resulting in a $(\mathbb{R}^4)^{20} = \mathbb{R}^{80}$ representation. Due to the construction of the pose vector, this descriptor is rotation invariant. The SMPL parameters were augmented with dynamic soft tissue deformation relative to each motion (called DMPL) and use to transform the original Dyna dataset to the DFAUST dataset, with better correspondance with the scan. They use for this goal much more information such as texture information from body videos, and

the shape vector is retrieved using gender information. We prefer comparing on Dyna dataset rather than DFAUST dataset, allowing us to compare faithfully to the SMPL body pose descriptor. In order to build the pose vectors, a costly fitting method is used along each sequence (accounting in minutes for a single shape). The pose vectors for 129 motions of Dyna where the fitting was successful, we added the SMPL Pose vector retrieved using available code https://github.com/vchoutas/smplx/ for the remaining 5 motions.

2 Comparison of SPD metrics for Gram-Hankel matrices

This section is dedicated to the comparison between Frobenius and Log Euclidean Riemannian Metric (LERM). The Gram-Hankel matrices are positive semidefinite matrices. Several metrics have been propose to compare positive semidefinite matrices. Table 1 shows the results of the comparison between Log Euclidean Riemannian Metric (LERM) and the Frobenius distance.

$$d_{LERM}(G_1, G_2) = ||\log(G_1) - \log(G_2)||_F,$$

where $\log(G) = P^T \log(\lambda) P$, where $G = P^T \lambda P$ is the eigen decomposition of the symmetric matrix G. We observe that the performance is lower than using

Representation	Gram-Hankel distance	Artificial dataset			Real dataset			Dyna dataset		
		NN	FT	ST	NN	FT	ST	NN	FT	ST
Current	Frobenius	100	100	100	92.5	66.0	78.5	59.0	34.1	50.4
	LERM	100	100	100	78.8	55.0	76.6	55.2	35.9	51.4
Absolute varifolds	Frobenius	100	100	100	95.0	66.6	80.7	60.4	40.0	55.9
	LERM	100	100	100	80.0	54.6	73.4	57.5	36.0	50.8
Oriented varifolds	Frobenius	100	100	100	93.8	65.4	78.2	60.4	40.8	55.9
	LERM	100	100	100	86.3	50.0	66.4	57.5	37.0	51.3

Table 1: Motion retrieval results for our approach with Log Euclidean Riemannian Metric (LERM). The results are displayed for CVSSP3D artificial and real datasets, and Dyna datasets

the Frobenius metric. This results confirms our choice of using Frobenius than LREM metric.

3 Extended discussion on the parameters r and σ

Effect of the sigma parameter. The performance relative to the σ parameter is displayed on the right of Figure 8 in the main paper for oriented varifolds on Dyna dataset. We observe first that the choice of σ has a significant impact on performance for NN and in the same time that the optimal σ for the NN is not the same as one for FT and ST, for a loss of around 2% in those metrics, which

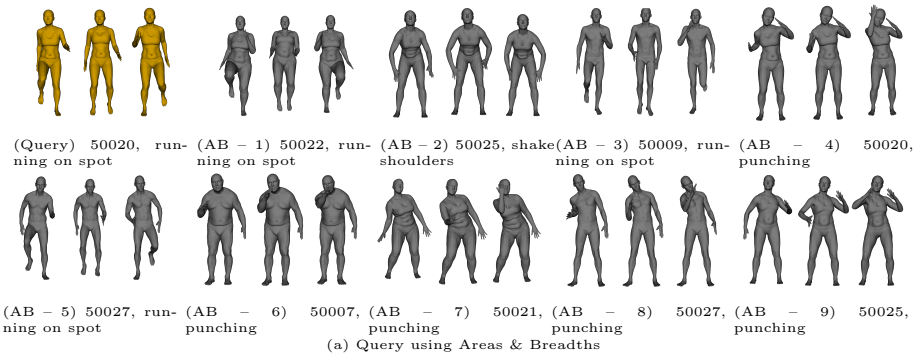
is less significant than the NN gain.
Effect of the choice of r. The performance relative to the r parameter is displayed on the left of Figure 8 for oriented varifolds on Dyna dataset. We observe first that the choice of r has a significant impact on performance and in the same time that the optimal r for the NN is not the same as one for FT and ST, for a loss of around 5% in those metrics.
Effect of normalizations. We present in Table 2 of the main paper the performances of oriented varifolds with the 2 normalization techniques presented here. The centroid normalization is essential to the good performance of our approach. In the mean time, the inner product normalization always implies significant boost for NN metric, but can induce a (non-significant) loss in ST and FT metrics.

4 Qualitative results: Queries on Dyna

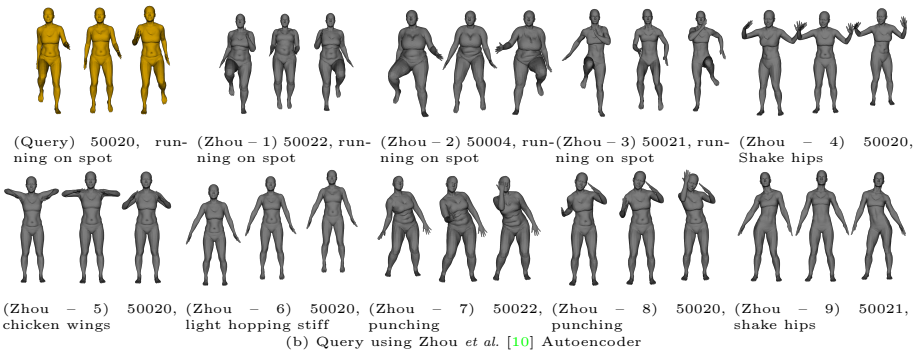
Figure 1 shows the results for SMPL, Zhou et al and Areas & Breadths. Although it is a the first tier is better for our approach in two manners: First we observe that there is no confusion between a motion and the motion of the same individual in our approach. Secondly, some drawbacks of the other methods appear: Areas & Breadths are symmetric descriptors and does not make the difference between a punching arm (from down to up) and the two arms that goes up and down when running, and we see a lot of punching motions retrieved (4 out of 6 wrong retrievals). Second, the autoencoder of Zhou *et al.*is not fully disentangled from the identity of the body and a lot of motions from the same identity are retrieved (4 out of 6 wrong retrievals). SMPL gives the best result, as expected from Table 1 of the paper. However, we observe also some sensitivity to the identity of the performer (2 out of 3 wrong retrievals).

5 Qualitative results on CVSSP3D real dataset.

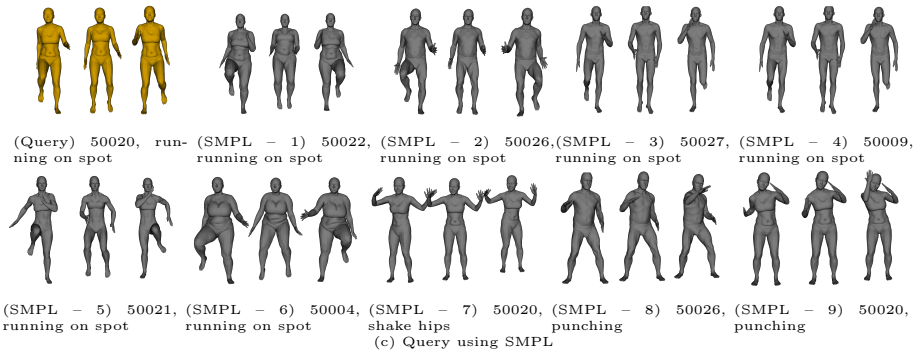
In the CVSSP3D real dataset, clothes worn by the subjects during the acquisition process induce topological and mesh noises (see Figure 1 and Figure 5(b) of the paper). The results on this dataset shows our method robustness to the noise and clothes present in clothed human dataset. The quantitative results in Table 1 (paper) show that our approach is robust to the noise and outperforms state-of-art methods on CVSSP3D real dataset in terms of NN. The confusion matrix of our approach (absolute varifolds) on CVSSP3D real dataset , in Figure 2 shows that our approach performs well on all human motions of the dataset. We display also a query with absolute varifolds, in Figure 3 (same query as the one displayed in [7]). Our approach is able to provide 6 out of the 7 walk motion, showing a slighlty better results compared to [7] (5 out of 7).



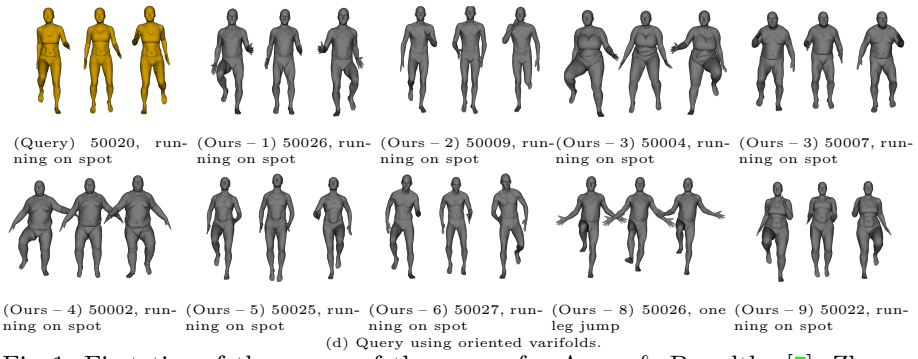
(a) Query using Areas & Breadths



(b) Query using Zhou *et al.* [10] Autoencoder



(c) Query using SMPL



(d) Query using oriented varifolds.

Fig. 1: First tier of the query of the paper for Areas & Breadths [7], Zhou *et al.* [10], SMPL [2] and oriented varifolds on the Dyna dataset. The query is in yellow and the results are sorted by closeness to the query using a given approach.

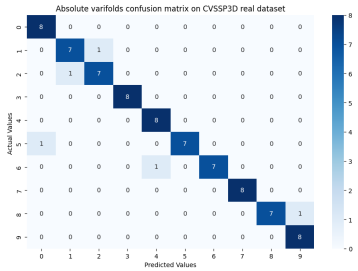
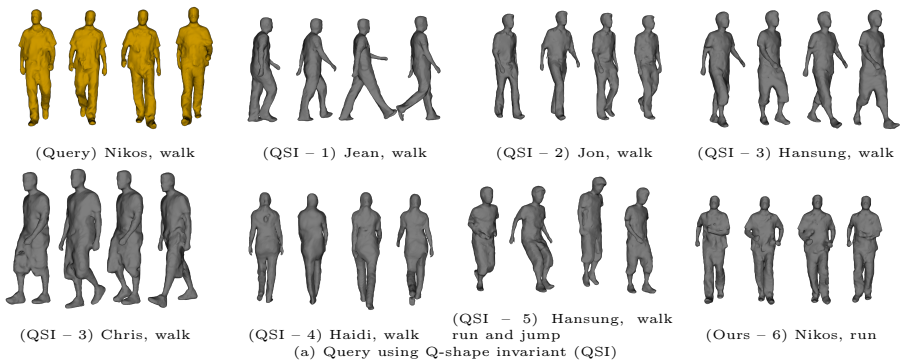


Fig. 2: NN Confusion matrix of absolute varifolds on CVSSP3D real dataset.



(a) Query using Q-shape invariant (QSI)

(b) Query using absolute varifolds

Fig. 3: First tier of the query of the paper for Q-shape invariant [7] and Absolute Varifolds. The query is in yellow and the results are sorted by closeness to the query using a given approach. The first query is directly taken from [7]. The query is in yellow and the results are sorted by closeness to the query using a given approach.

References

1. Aumentado-Armstrong, T., Tsogkas, S., Jepson, A., Dickinson, S.: Geometric disentanglement for generative latent shape models. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8180–8189 (2019) 1, 2
2. Bogo, F., Romero, J., Loper, M., Black, M.J.: FAUST: Dataset and evaluation for 3D mesh registration. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3794–3801 (2014) 2, 5
3. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Zafeiriou, S., Bronstein, M.M.: Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 7212–7221. IEEE (2019) 2
4. Cosmo, L., Norelli, A., Halimi, O., Kimmel, R., Rodolà, E.: Limp: Learning latent shape representations with metric preservation priors. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 19–35. Springer (2020) 2
5. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015) 2
6. Papadakis, P., Pratikakis, I., Theoharis, T., Passalis, G., Perantonis, S.: 3d object retrieval using an efficient and compact hybrid shape descriptor. In: Eurographics Workshop on 3D object retrieval (2008) 1
7. Pierson, E., Paiva, J.C.Á., Daoudi, M.: Projection-based classification of surfaces for 3d human mesh sequence retrieval. Computers & Graphics (2021) 1, 2, 4, 5, 6
8. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 4627–4635. IEEE Computer Society (2017) 2
9. Veinidis, C., Danelakis, A., Pratikakis, I., Theoharis, T.: Effective descriptors for human action retrieval from 3d mesh sequences. International Journal of Image and Graphics **19**(03), 1950018 (2019) 1
10. Zhou, K., Bhatnagar, B.L., Pons-Moll, G.: Unsupervised shape and pose disentanglement for 3D meshes. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 341–357. Cham (2020) 2, 5