Learning to Predict Anthropometric Landmarks via Feature Refinement

Yibo JIAO [1], Chang SHU [2], Dinesh K. PAI [1]
[1] University of British Columbia, Vancouver BC, Canada;
[2] National Research Council Canada, Canada

https://doi.org/10.15221/24.47

Abstract

Precise localization of anthropometric landmarks is essential in many applications --- including computer graphics, computer vision, biomechanics, and morphometric studies. Current methods in machine learning can produce dense correspondence with good global properties, but are poor at localizing specific landmarks. Landmarks are the most important correspondences for many applications and the only independently verifiable criteria. Here we propose a new method for localizing landmarks using learned features that significantly outperforms the state of the art.

Our method learns refined features that characterize the intrinsic and extrinsic geometry around the landmark, thereby making it easy to recognize and localize. We propose a novel loss function for training using two functions to evaluate the likelihood of a vertex being a target landmark: the landmark potential $(\mathcal{P})$, which characterizes the likelihood of a vertex being a landmark and is computed by the network from a given mesh, and the similarity function $(\mathcal{D})$, which measures the distance from a landmark, extrinsically or/and intrinsically, and provides supervision information for training. The network is trained to optimize the correlation between the two functions so that vertices that have high similarity with the target landmark also have high potential, and vice versa. In addition, we introduce nuclear-norm minimization to compute an optimal span of refined features for landmarks among training models. The resulting system is easy to implement and learns efficiently from a small number of meshes with identified landmarks.

Our implantation is available at https://github.com/yibojiao211/Learning_to_Predict_Landmarks.

Keywords: shape matching, deep learning, anthropometry

1. Introduction

Human shape correspondence is an essential task for many areas in computer vision and computer graphics, as well as in downstream applications ranging from video games to medicine. The problem is to establish a point-to-point mapping between scans or 3D meshes, of different persons or the same person in different poses. Once a correspondence is established, coordinates of points can serve as shape variables for statistical analysis and other downstream applications.

With the advent of 3D imaging technologies, it is now possible to digitize the human body at high resolution and obtain detailed shape information. To make full use of these data, the computer vision and computer graphics communities have focused on creating dense correspondence. Impressive results have been achieved [9,12,13,41]. This work, reviewed in the related work typically uses features based on shape or appearance data, but with no intrinsic importance other than their ability to facilitate the correspondence problem.

We focus on a special class of feature points called **landmarks**. As the colloquial usage of the word ‘landmarks’ suggests, they are important in their own right and used for guidance in the neighborhood of a landmark. Human anatomical landmarks are associated with deep human anatomical structures, and are essential in many applications, ranging from guiding surgery to dressing virtual clothing. More generally, in biology, they facilitate a statistical theory of shape [4,10,17]. In anthropometry, landmarks were studied extensively for measuring human body size and shape. As a result, precise definitions and standards for the characteristics and locations of landmarks have been developed [15].

An important type of landmark is associated with a location on a human bone (‘bony landmark’). In many individuals, these locations can only be found by human experts by palpating the bony protuberances on the human body prior to scanning, which is a tedious and expensive process. Due to these difficulties, only a small number of 3D scanning projects, notably, the CAESAR project [28], have included human landmarking. These datasets provide the ground truth validation for correspondence algorithms. In most cases, they serve as the only means of obtaining verifiable validation of a correspondence algorithm.

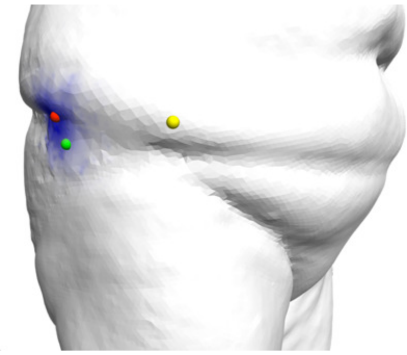

Figure 1 ASIS identified by different methods. Green: ground truth; Red: our method; Yellow: Deep Shells

For example, the Anterior Superior Iliac Spine (ASIS) is an important bony landmark at the hip that is difficult to identify based on intrinsic body shape alone, but is easy to identify by palpation. In Figure 1, the yellow dot depicts the location found by Deep Shells [12], a state-of-the-art method; it is far from the ground truth landmark, shown in green. As a preview, the prediction of our method, described in this paper, is shown in red. Thus, the challenge of accurately identifying these important points has remained unresolved in previous work.

In this paper, we propose a new method for identifying these important points, the anthropometric landmarks. Our objective is to train a network to predict landmark locations from 3D scans using the easily accessible geometric surface features. Our approach is based on refining features with deep neural networks to improve the accuracy of locating landmarks, an idea initially introduced by [20]. At a high level, we aim to refine the features such that all corresponding landmarks among training shapes have similar refined features, and points that are both extrinsically and intrinsically close to landmarks have similar refined features as landmarks. We propose two functions to evaluate the likelihood of a vertex being a target landmark. The potential function $\mathcal{P}$ measures the likelihood in feature spaces, which takes as a parameter the refined features produced by the network. The similarity function $\mathcal{D}$ measures the likelihood extrinsically or/and intrinsically, such as Euclidean or geodesic distances; it provides supervision information during training. We assume that landmark potential and similarity are correlated, thus the goal of the network is to optimize the correlation so that vertices that have high similarity with the target landmark also have high potential value and vice versa. The potential function automatically has a minimum value at the landmark; in this way, designing the function is meaningful. For any candidate human model, we can identify the vertex with the lowest potential as the landmark location.

2. Background and Related Work

2.1. Anthropometry

Anthropometry is the study of human body measurement. Traditionally, anthropometric data is collected through direct measurements using simple tools, such as rulers, calipers, and measuring tapes. Most measurements are defined based on the landmarks. Over the years, consensus and standards have been established regarding the locations of these landmarks [15].

With the advent of 3D scanning technology, many 3D anthropometric datasets have been developed. The CAESAR project [28] developed the first large-scale 3D anthropometric dataset, which contains 5,000 scans collected from individuals across North America and Europe. In addition to the 3D scans, it also included 73 landmarks hand-placed by experts, as well as traditional measurements. In academia, smaller but publicly available datasets, such as SCAPE [1] and FAUST [3]have been developed. These datasets provided dense point-to-point correspondence through fitting a template to the individual scans, which have served as benchmarks for many research works on human shape correspondence. In the FAUST dataset, painted markers and 17 manually placed landmarks were used to aid with registration, although the specific locations of these landmarks were not released.

Placing landmarks on the human body requires a strong understanding of human anatomy and practical experience, as it involves palpating the bony extrusions. Additionally, this process can be time-consuming. This prompted some early works on estimating landmarks on body scans [2,23,42]. These works focused either on easily identifiable landmarks or need much more work on individual analysis [23,35]. In [2], the problem is formulated as a probabilistic inference problem and a Markov network is trained to localize all 73 landmarks for the CAESAR scans. Reasonable results were achieved on a small test set, but the CAESAR training data contains small bumps at the landmark locations, which may influence the prediction.

2.2. 2D keypoint detection

Locating anthropometric landmarks in 3D shapes shares similarities with keypoint detection and pose estimation in 2D images. In 2D, keypoints can be any point of interest [8,22,33] with joint locations, facial landmarks[45,46] and hand landmarks [34] being particularly useful for tracking humans in videos. Early methods relied on the classification of salient image features. The literature on keypoint detection and pose estimation is extensive, but for our purposes, we will focus on the methods that leverage deep neural networks, which have shown great success.

Among the various approaches, regression-based methods have been particularly effective. Initially, these methods directly regressed the coordinates of the keypoints. However, more recent and successful techniques adopt a heatmap-based approach. Instead of directly predicting coordinates, they regress a heatmap, often using a Gaussian distribution centered on the keypoint [7,16,39,40]. For instance, Tompson et al. [39] constructed individual heatmaps for each joint and trained a network using a simple mean-square-error loss. Another notable approach proposed by Sun et al.[36] is integral regression, which employs an end-to-end differentiable objective function. These heatmap-based methods have proven to be robust and accurate for keypoint detection and pose estimation.

In our work, we adopt a similar strategy of training separate networks for each landmark. However, due to the nature of working with a 3D surface mesh, regressing a heatmap becomes challenging. Instead, we maximize the correlation coefficient between two functions that capture local surface features and position differences. This approach offers a simple yet efficient solution to address the task of landmark localization in meshes.

2.3. Localizing by dense correspondence

Localizing anthropometric landmarks can be approached by addressing a stronger problem -- dense shape correspondence. By obtaining a dense correspondence between two shapes and having the ground truth landmarks for one shape, we can, in principle, derive the landmark locations in the other shape.

Axiomatic methods for establishing point-to-point correspondence provide a baseline for landmark localization [11,24,29], with many of them being based on functional maps [25]. In recent works, learning techniques have been applied, resulting in improved accuracy [9,12,13,14,19,20,30,37]. Litany et al. [21], introduced the idea of training a deep neural network to refine conventional surface descriptors, such as the SHOT descriptor. More recently, Trappolini et al. [41] proposed a transformer-based method. However, finding dense correspondences can be computationally intensive, particularly for high-resolution scans. For learning-based methods, the challenge becomes even more pronounced when training and testing with high-resolution real human scans that often contain partiality and noise. Furthermore, while these dense correspondence methods have achieved good overall matching results, they may fall short when it comes to the most critical landmark locations. We demonstrate this in the result section.

Kovnatsky et al. [18] proposed a method for finding functional correspondence through matrix completion. This method leverages the parsimony principle in statistical inference and model selection [27], enforcing low rank and small norm to represent the correspondence matrix with a small number of degrees of freedom. In our work, we also employ the parsimony principle and devise a cost function based on the nuclear norm [6].

3. Method

3.1. Formulation

We label k training samples in total as $[M_1, M_2, \ldots, M_k]$, where M_i is a mesh comprising an array of n_i vertex positions V_i and a set of m_i faces F_i. M_i can be raw output of a scanner (e.g., after surfacing with marching cubes), but maybe noisy with holes, without a fixed topology. For the given mesh M_i, we first compute point-wise SHOT descriptors [38], S_i for each vertex. Let G_i be the input feature of M_i, and G_{i_j} be the feature of the j-th point of M_i. G_{i_j} is formulated by concatenating the S_{i_j} and 3D position of the point, V_{i_j}. K is the dimension of the SHOT descriptor, it can be represented as:

$$M_i \xrightarrow{pre-process} \begin{bmatrix} S_i \\ V_i \end{bmatrix} = G_i = \left[G_{i_1} \ldots G_{i_n}\right] \in \mathbb{R}^{(K+3)\times n}$$

We take this concatenated matrix as input of the network, in other words, the unrefined features. Let the parameters the neural network be θ,the refined features are:

$$G_i^* = f_\theta(G_i) = \left[G_{i_1}^* \dots G_{i_n}^*\right] = \left[f_\theta(G_{i_1}) \dots f_\theta(G_{i_n})\right]$$

Specifically, we represent unrefined features of the target landmark of M_i by $\mathcal{L}$, let the index of the target landmark be l for shape M_i,$G_{i_l} = \mathcal{L}_i$, then

$$\begin{aligned} G_i &= \left[G_{i_1} \dots G_{i_l} \dots G_{i_n}\right] \\ &= \left[G_{i_1} \dots \mathcal{L}_i \dots G_{i_n}\right] \end{aligned}$$

We define the landmark matrix $\mathcal{L}$ of the k training samples to be formulated by all unrefined features of landmarks of the k samples, then

$$\mathcal{L} = [\mathcal{L}_1 \dots \mathcal{L}_k]$$

Furthermore, we design our potential and similarity functions to be point-wise scalar values, then we optimize correlation. The pipeline of our training is shown below in Figure 2.

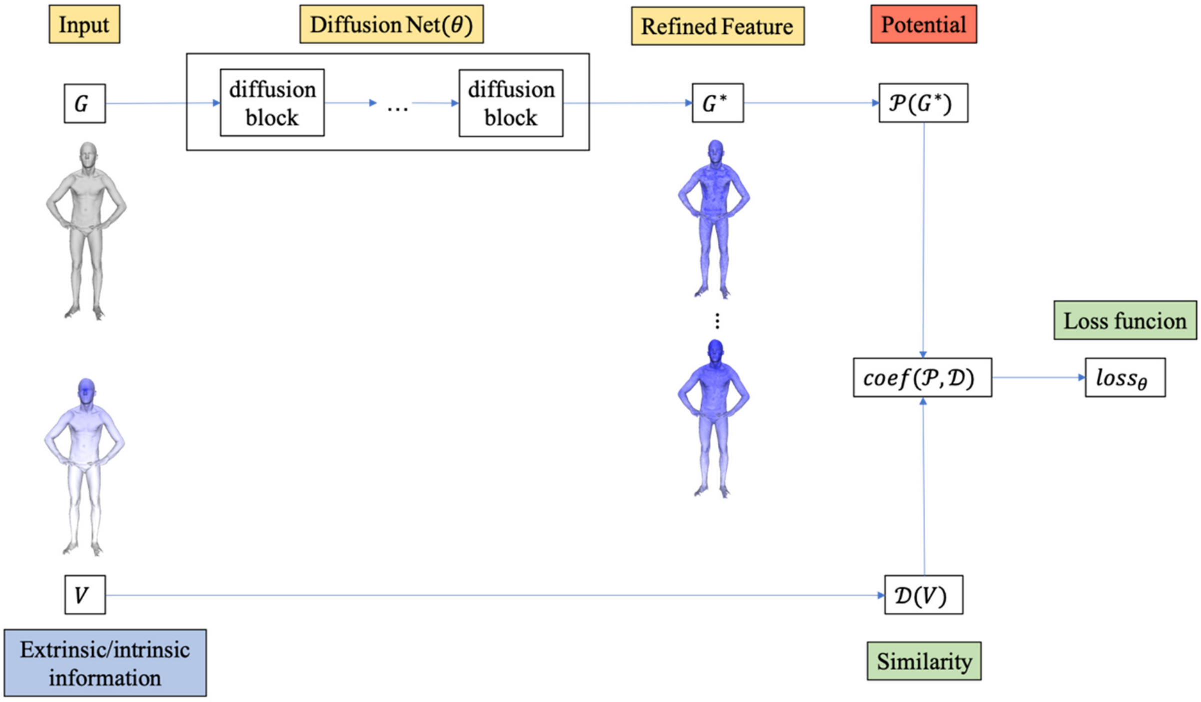

Figure 2 Overview of the general training framework. Input point-wise SHOT $G, \mathcal{L}$ and extrinsic information V, L from training meshes. $G, \mathcal{L}$ are passed through a neural network made of identical operation blocks [32] to get refined features $G^*, \mathcal{L}^*$, which are further used for computing potential functions. Meanwhile, extrinsic/intrinsic information produces the similarity functions. With point-wise potential and similarity values, the correlation between the two functions contributes to the loss function l_θ. User-defined components in our framework are operation blocks, potential, similarity, and loss functions, a simple example of choices is shown in section 4.3.

3.2. Potential and similarity functions

We define two functions to measure the likelihood of a vertex being the target landmark. The potential function $\mathcal{P}$ computes refined feature space likelihood and thus takes a parameter of refined features. The similarity function $\mathcal{D}$ computes extrinsic/intrinsic distances, which are provided in the training data and work as supervision labels. For testing samples, similarity function values are unknown, thus our goal is to extract refined features by learning. As a result, the learned features capture the nature of the geometry of input meshes and can be used to evaluate the landmark likelihood without further information. Our method trains a neural network so that potential and similarity functions are highly correlated.

Please note that due to the typical convention in machine learning to minimize functions, we define the potential and similarity functions with a sign opposite to that of common usage. That is, low values of "potential" mean a point is likely to be landmark and low values of "similarity" mean that a point is similar to a landmark.

3.3. Optimizing correlations and nuclear norms

We choose DiffusionNet [32] as a part of the learning architecture. The naive approach to feature refinement is to simply solve the problem by regression. The potential and similarity functions are simply point-wise ℓ^2-norm of refined feature and pre-computed Euclidean/geodesic distance to the target landmark, and we can use the MSE loss between the two functions. We initially experimented with this approach.

Instead, we choose the coefficient of correlation as the cost function because we are concerned with both the direction of change and the degree of consistency between the potential and similarity functions. This approach allows us to regress a heatmap without the need to set up a heatmap on the 3D surface, which is both difficult and inefficient. On the other hand, the naive way of minimizing the mean square error between the functions only takes into account their numerical differences. However, since the two functions are measured differently - one in the feature space and the other in Euclidean or geodesic distance - it is more effective to measure their relative direction of change rather than the pointwise numerical difference.

Recall that $\mathcal{L}^* = f_\theta(\mathcal{L}) = [\mathcal{L}_1^*, \dots, \mathcal{L}_k^*]$, which are all refined features of target landmark of k training samples. Nuclear-norm-based models were proposed to solve image restoration and denoising problems [43,44]. Prior works exploited low-rank properties of patch groups to capture similarities inside batches and injected low-rank priors to capture the nature of high dimensional data [31]. The low-rank model has been proven to be able to extract and analyze features that are linearly dependent and sparse [5]. In this work, we consider low-rank properties that exist in landmarks because of similar geometry, thus we optimize the nuclear norm of features. The goal of injecting low-rank priority is to train a network so that refined features of landmarks are similar for all shapes.

Since $\mathcal{L}^*$ gathers all refined descriptors from all training shapes, by injecting low-rank properties, we minimize the rank of $\mathcal{L}^*$,which is equivalent to minimizing ℓ^0-norm of singular values of $\mathcal{L}^*$, $\|\sigma(\mathcal{L}^*)\|_0$. In [6], it had been proven that ℓ^0-norm minimization can be solved by ℓ^1-relaxation, thus with ℓ^1-norm relaxation, we minimize nuclear norm of $\mathcal{L}^*$:

$$\min \operatorname{rank}(\mathcal{L}^*)$$

$$\Leftrightarrow \|\sigma(\mathcal{L}^*)\|_0$$

$$\Leftrightarrow \|\sigma(\mathcal{L}^*)\|_1 = \sum_{i=1}^{k} |\sigma_i(\mathcal{L}^*)|$$

$$\Leftrightarrow \min \|\mathcal{L}^*\|_* = \|f_\theta(\mathcal{L})\|_*$$

Relying on the geometrical similarity for landmarks, adding low-rank priors helps to find θ so that our network refines input features in a way that maximally captures the nature of similarity between landmark points among different shapes. However, simply optimizing the nuclear norm of $\mathcal{L}^*$ has a trivial solution of $\theta = 0$, we use nuclear norm minimization as a constraint to our loss objective. We can also add constraints to the above program so that the refined features are not only similar to raw input features but also have low-rank properties. Thus, we solve:

$$\lambda_* \|\mathcal{L}^*\|_* + \lambda_2 \|\mathcal{L}^* - \mathcal{L}\|_2^2$$

To evaluate point-wise potential, we simply compute the distance between the point to the target landmark, for input G_j and point i on the shape, the potential value is:

$$\mathcal{P}\left(G_{j_i}^*, \mathcal{L}_j^*\right) = \left\|\mathcal{L}_j^* - G_{j_i}^*\right\|_2^2 \qquad (1)$$

To involve both intrinsic and extrinsic information for similarity function, we compute weighted addition of geodesic distance or Euclidean distance and SHOT differences:

$$\mathcal{D}\left(G_{j_i}, \mathcal{L}_j\right) = c_{ex}\left\|V_{j_i} - L_j\right\|_2^2 + c_{in}\left\|S_{j_i} - \mathcal{L}_j\right\|_2^2 \qquad (2)$$

We combine correlation and low-rank priority to formulate the loss function:

$$loss_\theta = -\lambda_c coef(\mathcal{P}, \mathcal{D}) + \lambda_* \|\mathcal{L}^*\|_* + \lambda_2 \|\mathcal{L}^* - \mathcal{L}\|_2^2 \qquad (3)$$

This formulation helps to maximize the correlation between point-wise potential and similarity function values and adds low-rank priors at the same time. In this way, we optimize θ so that (a) the refined features of landmarks span low-dimensional subspaces and (b) non-landmark features differ from landmark features in a way that is close to extrinsic distances.

3.4. Pipeline Overview

For training, we used the CEASAR dataset which contains landmark labels. We used Equation (1) only in training phase and $\mathcal{L}^*$ and G^* are refined features of landmarks and all points. At testing time, only refined features of all points were used. These features are trained to maximize the correlation function such that points close to the training landmark have small potential values and points far away from the landmark have large potential values. When testing, the refined features also have this property. This is equivalent to have a constant $\mathcal{L}^*$ in Equation (1). Therefore, we can simply compute the point-wise potential value as the norm of the refined feature. The point with the minimal potential value is the predicated landmark on each testing shape.

4. Results

4.1. Implementation Details

We used 4 blocks of DiffusionNet in the network to compute potential. The input dimension of features is 355 (352 for SHOT and 3 for position); the output can be varied. In order to add the ℓ^2-norm regularization term (the last term of Equation 3), the output dimension of DifussionNet must also be 355; adding this term is optional. To reduce the computational burden of the training process we chose the output dimension to be 8. This could be reduced further; we found that even making the output dimension to be 1 gives very good results. We choose $\lambda_c = 0.95$, and other weights to be summed to 0.05. We trained using 50 shapes of removed FAUST scan rather than FAUST registration, each shape has about 120K vertices. Our method is efficient to predict and train with such high-resolution shapes. We will provide code with the published paper to generate the results in the paper. The provided code can run with CPU.

4.2. Landmark potential

For testing the network, we computed the point-wise potential values for different landmarks. Figure3 provides a visualization of the potential over a scan, with color-coded values. Point estimates are computed by taking the mean of points with the top 1% of the values (colored blue in the close-up in the second row) and shown in red. Unlike correspondence-based localization, we have evaluated all vertices on the mesh. For geometry factor-dominated landmarks, our method localizes the prediction to a very small area, thus producing a very precise result. For landmarks that require palpation to measure, the area of localization is relatively large, and the minima of potential functions might be slightly far away from the ground truth. However, for such landmarks, our method still outperforms the methods of shape matching, and has a reasonable amount of error. Furthermore, one can use the potential distribution to improve the result because outliers are easily detected. In other words, unreasonable low potential values can be easily filtered, and using barycentric coordinates with points of low potential values instead of returning the minima can help to improve the localization.

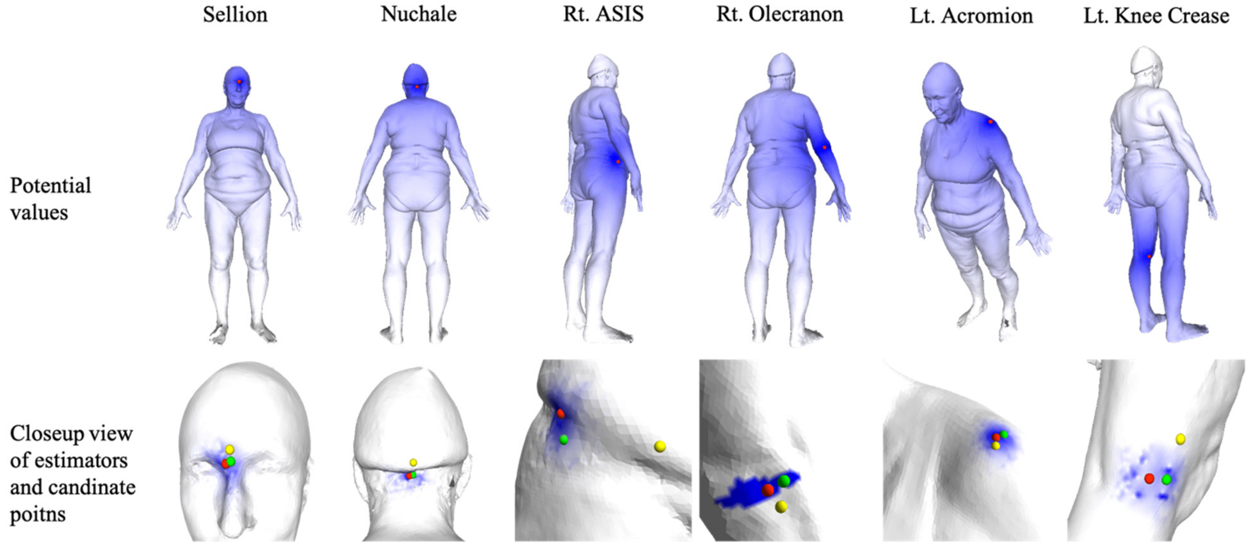

Figure 3 Visualization of point-wise potential values of an example shape. The first row contains color-coded shapes for potential values and red spheres as our predictions. The second row contains closeup views of landmark areas, the top candidate points are colored blue. Red: our results, green: ground truths, yellow: Deep Shell correspondences

4.3. Robustness to pose change

We evaluated the point estimates of different landmarks on in different poses that were not used for training. Figure 4 shows scans in four different poses of the same subject. The landmarks are quite robust even with holes in the scans.

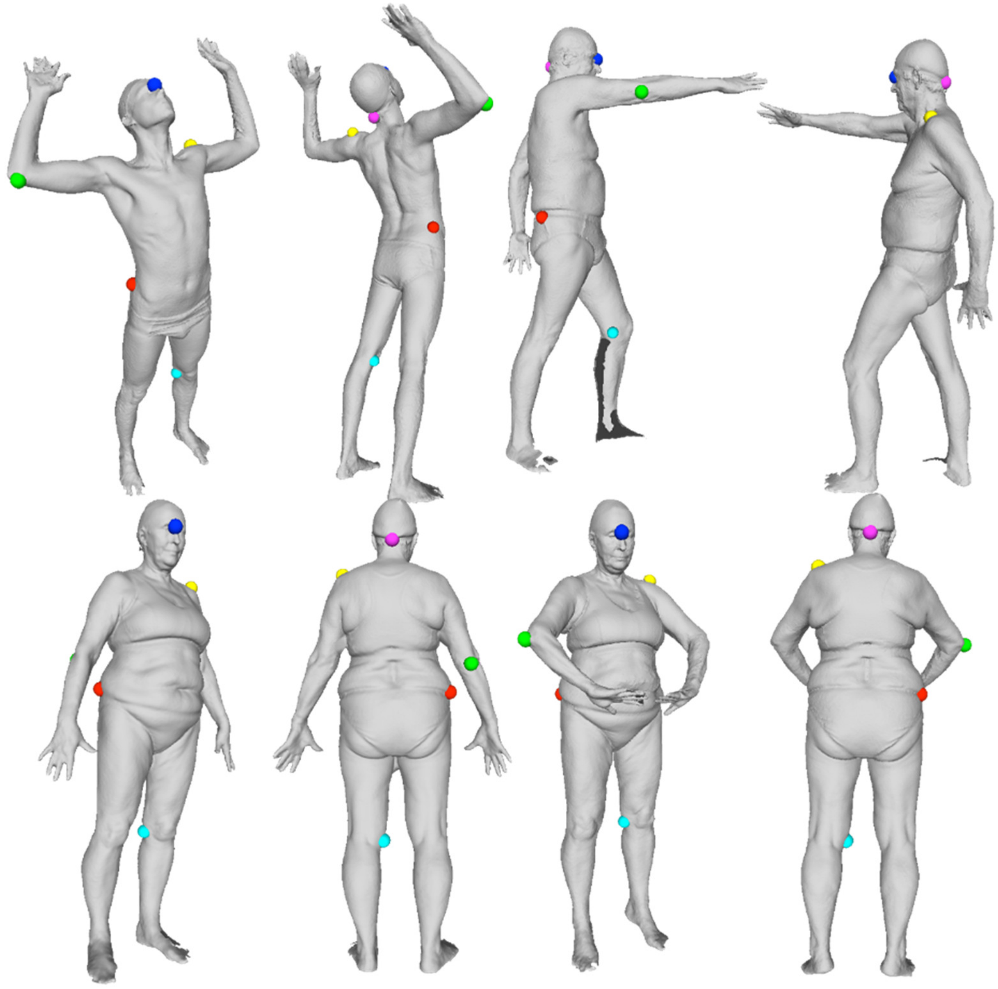

Figure 4 Estimators of predicted landmarks on shapes of various poses. Blue: Sellion, green: Rt. Olecranon, yellow: Lt. Acromion, red: Rt. ASIS, cyan: Lt. Knee Crease, magenta: Nuchale

4.4. Localization accuracy

Early learning-based methods were supervised, requiring extensive amounts of dense 3-D labelling data and presenting substantial challenges. Consequently, recent efforts have predominantly concentrated on unsupervised approaches, which have yielded better results compared to supervised alternatives. For instance, DeepShells [12] outperforms 3D-CODED [13], FMNet [20], and GeoFMNet [9].

For each of the target landmarks, we trained the network separately and recorded the error of localization. The error was compared to that of landmarks predicted by Deep Shells, a representative of the current state-of-the-art methods for computing dense correspondences. The results are shown in Table 1. Our method has a smaller error for almost all landmarks, typically by a large margin.

Table 1. Error of our method and Deep Shells. All values are average errors among all 70 testing samples for both methods. DS stands for Deep Shells for finding point-wise correspondences, units of mm.

Landmark	DS	Ours	Landmark	DS	Ours
1 Sellion	11.5	3.05	38 Rt. Dactylion	24.5	2.87
2 Rt. Infraobital	11.2	4.56	39 Rt. Ulna Styloid	10.6	5.83
3 Lt. Infraobital	11.9	2.08	40 Rt. Metacarpal Phal. V	15.1	5.68
4 Supramenton	9.86	1.89	41 Lt. Acromion	6.93	4.19
5 Rt. Tragion	7.29	2.41	42 Lt. Axilla, Ant	8.01	6.83
6 Rt. Gonion	7.33	3.42	43 Lt. Radial Styloid	24.2	2.85
7 Lt. Tragion	8.86	4.19	44 Lt. Axilla, Post	12.5	10.4
8 Lt. Gonion	4.90	2.13	45 Lt. Olecranon	15.7	6.42
9 Nuchale	2.96	4.18	46 Lt. Humeral Lateral Epicn	14.1	7.41
10 Rt. Clavicle	2.73	3.36	47 Lt. Humeral Medial Epicn	15.2	4.64
11 Suprasternale	9.13	8.02	48 Lt. Radiale	14.7	7.49
12 Lt. Clavicle	4.31	2.85	49 Lt. Metacarpal Phal. II	22.4	15.6
13 Rt. Thelion	7.33	2.37	50 Lt. Dactylion	80.5	2.55
14 Lt. Thelion	8.45	4.26	51 Lt. Ulnar Styloid	23.8	6.35
15 Substernal	10.6	8.76	52 Lt. Metacarpal Phal. V	22.1	3.61
16 Rt. 10th Rib	18.1	6.97	53 Rt. Knee Crease	4.71	5.95
17 Rt. ASIS	19.8	11.5	54 Rt. Femoral Lateral Epicn	4.75	4.45
18 Lt. 10th Rib	15.4	8.39	55 Rt. Femoral Medial Epicn	6.64	5.68
19 Lt. ASIS	13.9	9.76	56 Rt. Metatarsal Phal. V	8.95	4.67
20 Rt. Iliocristale	11.1	9.48	57 Rt. Lateral Malleolus	9.82	6.87
21 Rt. Trochanterion	14.4	10.9	58 Rt. Medial Malleolus	11.9	6.56
22 Lt. Iliocristale	11.4	8.76	59 Rt. Sphyrion	91.6	5.73
23 Lt. Trochanterion	10.4	10.2	60 Rt. Metatarsal Phal. I	11.4	9.16
24 Cervicale	7.69	5.33	61 Rt. Calcaneous, Post	9.44	9.39
25 10th Rib Midspine	10.8	9.38	62 Re. Digit II	11.5	10.7
26 Rt. PSIS	8.93	8.14	63 Lt. Knee Crease	4.67	2.52
27 Lt. PSIS	9.02	8.27	64 Lt. Femoral Lateral Epicn	6.74	4.38
28 Waist	12.9	6.54	65 Lt. Femoral Medial Epicn	7.16	3.84
29 Rt. Acromion	21.9	4.55	66 Lt. Metatarsal Phal. V	16.1	3.73
30 Rt. Axilla, Ant	10.2	6.87	67 Lt. Lateral Malleolus	9.48	5.11
31 Rt. Radical Styloid	11.1	2.09	68 Lt. Medial Malleolus	9.09	6.59
32 Rt. Axilla, Post	16.3	9.35	69 Lt. Sphyrion	8.96	5.73
33 Rt. Olecranon	10.2	8.88	70 Lt. Metatarsal Phal. I	4.49	4.28
34 Rt. Humeral Lateral Epicn	8.58	3.44	71 Lt. Calcaneous, Post	9.46	7.89
35 Rt. Humeral Medial Epicn	12.5	4.26	72 Lt. Digit II	11.9	2.96
36 Rt. Radiale	8.73	7.79	73 Crotch	16.2	7.82
37 Rt. Metacarpal Phal. II	11.1	10.9			

To evaluate and compare our proposed method and other state-of-art methods that compute dense shape correspondences we compared FMNet [20], Deep Shells [12], and transmatching [41]. The results are shown in Table 2. In this study, we trained with 50 FAUST scan shapes and tested with 50 FAUST scans and 50 CAESAR scan shapes. We noticed that there are small bumps on the CAESAR meshes that might affect our learning process, therefore we add an extra step of remeshing to the pre-processing operation in order to remove any factors by bumps of the scans and indices of vertices. We used the isotropic explicit remeshing algorithm proposed in [26]. We selected 6 landmarks that are easy to hand-place on FAUST scans and important for template fitting in order to compare all methods.

Table 2. We compare FMNet, Deep Shells, transmatching, and our method by training with 50 FAUST scan shapes, and testing with 50 FAUST scans that were not used for training (labeled `FS') and 50 CAESAR scans (labeled `C'). All shapes are normalized to having $2m^2$ surface area for male shapes and $1.8m^2$ for female shapes. In this way the unit of errors in the table is meter.

Methods		Fmnet		Deep Shells		transmatching		Ours	
Train on FAUST scan		Test on							
Landmarks		FS	C	FS	C	FS	C	FS	C
1	Sellion	0.075	0.138	0.013	0.025	0.035	0.127	0.007	0.004
41	Lt. Acromion	0.078	0.146	0.017	0.102	0.091	0.329	0.011	0.027
63	Lt. Knee Crease	0.064	0.187	0.049	0.056	0.051	0.139	0.012	0.012
9	Nuchale	0.058	0.107	0.032	0.011	0.027	0.232	0.008	0.023
16	Rt. ASIS	0.086	0.212	0.027	0.087	0.082	0.261	0.014	0.034
33	Rt. Olecranon	0.052	0.158	0.026	0.064	0.095	0.366	0.025	0.009

For the 3 dense shape correspondence methods, the prediction of localization is computed by using one template with ground truths of landmarks (for the CAESAR dataset, we used the included landmark information, for FAUST scan data, landmarks are hand-placed on FAUST registration, and we use the registration to obtain landmarks on FAUST scans for all shapes.). Dense correspondences of the template shape and test shapes are then computed. For each pair of template and test shape, we first obtain all points that are within the $2cm^2$ area of landmarks on the template and their matching points. We then use the matching points on the test shape to compute the average distance to the landmark on the test shapes. For our method, we simply select the top 1% points of the highest potential and calculate the distance between the average point and the landmark ground truth. Table 2 shows that our method outperforms all the other methods, in many cases by an order of magnitude.

5. Conclusion

We proposed a new method based on feature refinement for predicting anthropometric landmarks. Our method handles high-resolution real scans of human bodies, that may suffer from partial data and noise. Instead of computing point-wise correspondences, our method takes advantage of the architecture of the DiffusionNet and refines SHOT descriptors to predict target landmarks of choice. We show that our method outperforms other state-of-the-art methods for high-resolution real scans and is able to localize landmarks for inter-dataset shapes.

We proposed a novel loss function for training, using two functions to evaluate the likelihood of a vertex being a target landmark: a landmark potential that is learned and used for identifying the landmark, and a similarity function that provides supervision information for training. We also introduced a nuclear norm term with regularization to find low-rank solutions. Our results show that this approach is highly effective, and can learn from a small amount of data (e.g., 50 scans with landmark information).

Even though we focused on identifying these important landmark points in this paper, our method may also benefit from finding dense correspondence between human bodies by providing an additional source of information derived from bony landmarks identified by experts. We leave this for future work. In general, our results develop a bridge between recent machine learning methods for computing correspondences and the wealth of knowledge and applications of classical anthropometric methods.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. In ACM SIGGRAPH 2005, pages 408–416, 2005.

[2] Z. Ben Azouz, C. Shu, and A. Mantel. Automatic locating of anthropometric landmarks on 3D human models. In Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), pages 750–757, 2006. doi: 10.1109/3DPVT.2006.34.

[3] F. Bogo, J. Romero, M. Loper, and M. J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3794–3801, 2014.

[4] F. L. Bookstein. Morphometric Tools for Landmark Data. Cambridge University Press, 1991.

[5] T. Bouwmans and E. H. Zahzah. Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance. Computer Vision and Image Understanding, 122:22–34, 2014. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2013.11.009.

[6] E. Candes and J. Romberg. l1-magic: Recovery of sparse signals via convex programming. URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf, 4(14):16, 2005.

[7] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In Proceedings of the IEEE international conference on computer vision, pages 1212–1221, 2017.

[8] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 224–236, 2018.

[9] N. Donati, A. Sharma, and M. Ovsjanikov. Deep geometric functional maps: Robust feature learning for shape correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8592–8601, 2020.

[10] I. Dryden and K. Mardia. Statistical Shape Analysis. Wiley, 2002.

[11] M. Eisenberger, Z. Lahner, and D. Cremers. Smooth shells: Multi-scale shape registration with functional maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12265–12274, 2020.

[12] M. Eisenberger, A. Toker, L. Leal-Taix., and D. Cremers. Deep shells: Unsupervised shape correspondence with optimal transport. Advances in Neural Information Processing Systems, 34, 2020.

[13] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. 3D-CODED: 3D correspon325 dences by deep deformation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 230–246, 2018.

[14] O. Halimi, O. Litany, E. R. Rodol., A. M. Bronstein, and R. Kimmel. Unsupervised learning of dense shape correspondence. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4365–4374, 2019.

[15] ISO 7250-1:2017. Basic human body measurements for technological design—Part 1: Body measurement definitions and landmarks. Standard, International Organization for Standardization, Aug. 2017.

[16] S. ke, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. pages 5686–5696, 06 2019. doi: 10.1109/CVPR.2019.00584.

[17] D. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. Bulletin of the London Mathematical Society, 16:81–121, 1984.

[18] A. Kovnatsky, M. M. Bronstein, X. Bresson, and P. Vandergheynst. Functional correspondence by matrix completion. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 905–914, 2015.

[19] L. Li, N. Donati, and M. Ovsjanikov. Learning multi-resolution functional maps with spectral attention for robust shape matching. In Advances in Neural Information Processing Systems, 2022.

[20] O. Litany, T. Remez, E. Rodola, A. Bronstein, and M. Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[21] O. Litany, E. Rodol., A. M. Bronstein, and M. M. Bronstein. Fully spectral partial shape matching. In Computer Graphics Forum, volume 36, pages 247–258. Wiley Online Library, 2017.

[22] G. Loy and A. Zelinsky. Fast radial symmetry for detecting points of interest. IEEE Transactions on pattern analysis and machine intelligence, 25(8):959–973, 2003.

[23] S. Luo, Q. Zhang, and J. Feng. Automatic location and semantic labeling of landmarks on 3D human body models. Computational Visual Media, pages 1–18, 2022.

[24] S. Melzi, J. Ren, E. Rodol., A. Sharma, P. Wonka, and M. Ovsjanikov. Zoomout: spectral upsampling for efficient shape correspondence. ACM Trans. Graph., 38(6):155:1–155:14, 2019.

[25] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: a flexible representation of maps between shapes. ACM Transactions on Graphics (ToG), 31(4): 1–11, 2012.

[26] N. Pietroni, M. Tarini, and P. Cignoni. Almost isometric mesh parameterization through abstract domains. IEEE Transactions on Visualization and Computer Graphics, 16(4):621–635, 2009.

[27] J. Rissanen. Modeling by shortest data description. Automatica, 14(5):465–471, 1978.

[28] K. Robinette, H. Daanen, and E. Paquet. The caesar project: a 3-d surface anthropometry survey. In Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062), pages 380–386, 1999. doi: 10.1109/IM.1999.805368.

[29] E. Rodol., L. Cosmo, M. M. Bronstein, A. Torsello, and D. Cremers. Partial functional correspondence. In Computer graphics forum, volume 36, pages 222–236. Wiley Online Library, 2017.

[30] J.-M. Roufosse, A. Sharma, and M. Ovsjanikov. Unsupervised deep learning for structured shape matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1617–1627, 2019.

[31] T. Sakai. Unsupervised deep learning by injecting low-rank and sparse priors. 2021. doi: 10.48550/ARXIV.2106.10923. URL https://arxiv.org/abs/2106.10923.

[32] N. Sharp, S. Attaiki, K. Crane, and M. Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. ACM Trans. Graph., 41(3), mar 2022. ISSN 0730-0301. doi: 10.1145/ 3507905. URL https://doi.org/10.1145/3507905.

[33] Z. Shu, S. Xin, X. Xu, L. Liu, and L. Kavan. Detecting 3D points of interest using multiple features and stacked auto-encoder. IEEE transactions on visualization and computer graphics, 25(8):2583–2596, 2018.

[34] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1145–1153, 2017.

[35] R. Sitnik and M. Witkowski. Locating and tracing of anatomical landmarks based on full-field four-dimensional measurement of human body surface. Journal of Biomedical Optics, 13(4): 044039, 2008. doi: 10.1117/1.2960017. URL https://doi.org/10.1117/1.2960017.

[36] X. Sun, B. Xiao, F. 385 Wei, S. Liang, and Y. Wei. Integral human pose regression. In Proceedings of the European conference on computer vision (ECCV), pages 529–545, 2018.

[37] R. S. Sundararaman, R. Marin, E. Rodola, and M. Ovsjanikov. Reduced representation of deformation fields for effective non-rigid shape matching. Advances in Neural Information Processing Systems, 35:10405–10420, 2022.

[38] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. ECCV'10, page 356–369, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 364215557X.

[39] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[40] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.

[41] G. Trappolini, L. Cosmo, L. Moschella, R. Marin, S. Melzi, and E. Rodol.. Shape registration in the time of transformers, 2021.

[42] S. Wuhrer, Z. B. Azouz, and C. Shu. Posture invariant surface description and feature extraction. In IEEE Conference on Computer Vision and Pattern Recognition, pages 374–381, 2010.

[43] N. Yair and T. Michaeli. Multi-scale weighted nuclear norm image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[44] C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, and Y. Zhang. Depth image denoising using nuclear norm and learning graph model. ACM Trans. Multimedia Comput. Commun. Appl., 16(4), dec 2020. ISSN 1551-6857. doi: 10.1145/3404374. URL https://doi.org/10.1145/3404374.

[45] S. Yin, S.Wang, X. Chen, E. Chen, and C. Liang. Attentive one-dimensional heatmap regression for facial landmark detection and tracking. In Proceedings of the 28th ACM International Conference on Multimedia, pages 538–546, 2020.

[46] S. Zulqarnain Gilani, F. Shafait, and A. Mian. Shape-based automatic detection of a large number of 3D facial landmarks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4639–4648, 2015.