# A Framework for an Accurate Point Cloud Based Registration of Full 3D Human Body Scans*

Vladislav Golyanik[1,2] Gerd Reis[2] Bertram Taetz[1] Didier Stricker[1,2]
[1]Department of Computer Science, University of Kaiserslautern
[2]Department Augmented Vision, DFKI GmbH

## Abstract

*Alignment of 3D human body scans is a challenging problem in computer vision with various applications. While being extensively studied for the mesh-based case, it is still involved if scans lack topology. In this paper, we propose a practical solution to the point cloud based registration of 3D human scans and a 3D human template. We adopt recent advances in point set registration with prior matches and design a fully automated registration framework. Our framework consists of several steps including establishment of prior matches, alignment of point clouds into a common reference frame, global non-rigid registration, partial non-rigid registration, and a post-processing step. We can handle large point clouds with significant variations in appearance automatically and achieve high registration accuracy which is shown experimentally. Finally, we demonstrate a pipeline for treatment of social pathologies with animatable virtual avatars as an exemplary real-world application of the new framework.*

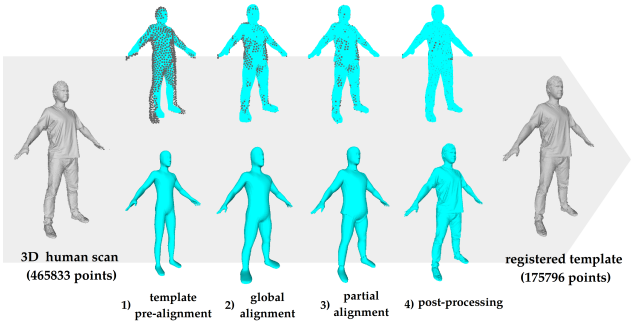


**Figure 1:** An overview of the proposed framework for a point cloud based 3D human scan-template registration with prior matches. On the left: the input 3D scan. The goal is to align it with a pre-defined full-body template. On the right: registration result. Middle: the scan overlayed with results of individual steps (top row) and the sequence of the deforming template progressing through the pipeline (bottom row).

## 1 Introduction

Alignment of real-world 3D human scans is a challenging problem in computer vision. Our goal is to align or register a given raw 3D scan with a 3D body template, i.e., to recover correspondences and a displacement field of a 3D human template to a reference 3D human body scan. Accurate human body registration has various applications in statistical shape analysis [1], anthropometric measurement extraction [2] and rehabilitation, to name a few.

If a human body scan and a template are represented by meshes, i.e., polygonal networks with normals and defined point topology, the problem is specified as mesh registration. This is a well-studied area with many works covering articulated human body registrations [3]. If both a human scan and a template are represented by point clouds, i.e., bare 3D point coordinates, the problem remains generally unsolved, and there are only a few attempts to tackle this problem [4]. There is, although, a demand to register human body scans as point clouds. For instance, 3D reconstructions obtained on a multi-view system represent noisy point clouds, with structured outliers caused by variations in cloths and appearances. Moreover, in many real-world scenarios meshes might not be available.

Recent advances in rigid and non-rigid point set registration allow closing the gap mentioned above, which is our main contribution in this paper. Extended Coherent Point Drift (ECPD) [5, 6] — a state of the art probabilistic approach for point set registration — allows embedding prior matches into a registration procedure. It was shown that in cases when prior matches are available or can be established automatically, ECPD can handle articulated cases more accurately compared to CPD [5]. Thus, we adopt ECPD for point cloud based registration of human body scans and design a multiple-stage human body registration pipeline schematically shown in Fig. 1. To align a scan and a template into a common reference frame and to estimate scale, our pipeline contains a pre-alignment step. The pre-alignment is followed by a global non-rigid registration providing initialisation for partial non-rigid registration. On several stages of the registration pipeline, automatically established prior matches guide the registration procedure. At the same time, our method is semi-automatic, and neither depends on large data sets (in contrast to [1]) nor requires point topology.

## 2 Related work

Among related works, there are several approaches similar to ours. For an exhaustive overview of rigid and non-rigid point set registration algorithms, an interested reader may refer to [7, 5, 6]. Additionally, several works on general point set registration can be mentioned which are not covered by these papers. Eckart *et al.* [8] proposed a rigid point set registration method based on decoupling of Gaussian mixture model parameters combined with a faster and more robust optimisation over a resulting compact representation. A

*The work was partially funded by the EU 7th Framework Programme project AlterEgo (600610) and the national BMBF project DYNAMICS (01IW15003). The authors thank Norbert Schmitz and José Henriques for helping to record the data and create visualisations, Vladimir Hasko for the human body template design as well as several anonymous volunteers for providing 3D body reconstructions used in the experiments. The contact e-mail address is *<first name>.<last name>@dfki.de*.

multiply-linked Gravitational Approach [9] allows registration of noisy point clouds with improved accuracy compared to other well parallelizable rigid point set registration methods, e.g., variants of Iterative Closest Point (ICP) [10], and more accurate but not fully parallelizable probabilistic methods.

Alignment of point sets involving complex displacement fields (with rigid and non-rigid components) constitutes the class of *articulated* point set registration methods. One of the early attempts was a generalisation of ICP [10] to articulated motion [11]. Especially for the case of articulated registration of 3D human body scans represented by point clouds, several approaches were recently proposed [12, 4]. These works offer efficient algorithms for registrations of the same bodies (or bodies very similar in appearance) in different poses. In contrast, we solve the problem of registering *dissimilar* human scans and a pre-defined template, with the goal of high accuracy, especially in the head and facial areas. In our case, 3D human scans can evince significant variation in clothes, hairstyle, body shape, proportions of individual body parts; we assume that poses of a scan and a template do not differ largely.

Dey *et al.* [13] proposed a markerless technique to align a human body mesh template to a new pose specified by a noisy point cloud of a human body. Similar to our approach, the method relies on body landmarks (head, hands and feet). However, we work with point clouds and do not solve the problem of pose alignment. The idea of alignment of a human body scan and a pre-defined template is used for anthropometric measurement extraction. Thus, in [14], a human scan and a template meshes are registered so that measurements can be consistently extracted on the template. A similar idea leading to high accuracy is adopted in [2], wherein a human body scan and a template are represented by point clouds. The method adopts global full body alignment, which is less accurate than our approach (though the accuracy suffices for the purpose of anthropometric measurement extraction). On the contrary, we use a segmented template, apply partial non-rigid registration for higher accuracy in regions with high structural variation and an optional post-processing.

## 3 The framework

In this section, we describe in detail the proposed framework for registration of full 3D human body scans.

### 3.1 A human body template

The proposed approach relies on a human body template which represents a segmented point cloud in a shape of a human body, see Fig. 2. The template was created by a designer and contains $1.75 \cdot 10^5$ points. Optionally, point topology and joints are available for it (Fig. 2-(b)). For an enhanced precision, our pipeline accepts prior matches — facial and body landmarks, if available (Fig. 2-(c), -(d)).

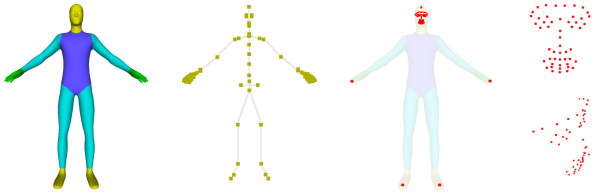


**Figure 2:** A full-body human template: (a) template with the pre-segmented body parts shown in different colours; (b) the skeleton of the template (can be optionally provided for the registration); (c) positions of the predefined facial and body landmarks on the template; (d) zoomed in facial landmarks in a frontal and a side view.

### 3.2 Overview of the framework

Assume we are given two 3D point clouds: the template $\mathbf{T}_{N\times 3} = (\mathbf{t}_1, \ldots, \mathbf{t}_N)^\mathsf{T}$ and the reconstruction $\mathbf{R}_{M\times 3} = (\mathbf{r}_1, \ldots, \mathbf{r}_M)^\mathsf{T}$ (reference). In ECPD, prior matches are provided by a set of indices $N_c \subset \mathbb{N}^2$ with an uncertainty parameter $\alpha$ (see Sec. 3.3 for details on establishment of prior matches). In the rigid case, ECPD [6] outputs parameters of rigid body motion and scaling to pre-align the template with a scan. In the non-rigid case, ECPD [5] outputs a displacement field aligning template and a reference as well as the probability of correspondences $P$. Further algorithmic details can be found in [5, 6].

Thus, the first step of the framework — *template pre-alignment* — consists in the alignment using rigid ECPD. Note that the joints of the template rigging are included into the registration as additional points. The second step — *global alignment* — consists in a rough non-rigid registration using non-rigid ECPD on the whole template (without segmentation). As can be observed in Fig. 1, this step reasonably accounts for overall body shape variations to obtain preliminary correspondences between the body parts of the template $\mathbf{T}$ (see Fig. 2) and the corresponding body parts of the reconstruction $\mathbf{R}$ using the correspondences of the highest probabilities, i.e., the nearest neighbours. Furthermore, due to the coherency constraint of the ECPD, the joints are automatically placed to the corresponding position w.r.t. the registered template. The correspondences at the surface of the template are subsequently used in the third step for the *partial alignment* of the corresponding parts of the scan.

The partial alignment is much more accurate, in particular for all extremal parts of the body and the hand regions (see Sec. 3.5 for resolution of challenging cases). Due to the coherency constraint, the registered surfaces are still rather smooth. In step four — the *post-processing* — the proximity of the noise-free registered template and the initial reconstruction is used to recover fine details via the projection techniques explained in Sec. 3.4.

### 3.3 Landmark Extraction

Suppose a reconstructed scan in an arbitrary pose and original images of the reconstruction are given. Since reconstructions can be noisy, it is hard to extract semantics from the scan (determine positions of the body parts). Thus, we detect faces in the input images using a face detector. We use the approach similar

to the one described in [5]. Once a suited image has been identified, the Chehra [15] facial feature detector is applied to the image. The detected feature points are transferred onto the scan by the projection to the corresponding nearest vertexes, since we know the optical relation between image and scan from the registration necessary to compute the reconstruction. The facial features robustly define a semantic coordinate system, i.e., the difference vector of the eye positions is sufficient to decide between left and right with respect to the reconstruction, and the difference vector between an eye and the mouth is sufficient to decide between up and down direction.

Next, the visual hull of the scan is voxelized. The scan is rendered using orthogonal projection from the six principal directions (imposed by the semantic vectors[1]). The voxelized version of the reconstruction is then thinned to the medial lines using a highly optimized implementation of [16]. This algorithm always reduces the volume to one-dimensional medial structures in contrast to other methods (e.g., [17]) which also produce two-dimensional medial structures (medial surfaces). Furthermore, the algorithm is pleasingly robust to noise and does not generate overly many branches. Once the 8-connected thinning is computed, the resulting voxelized representation of the medial structure is transformed into a graph which is in turn transformed into a minimum spanning tree using Kruskals algorithm [18], see Fig. 3-(a). Since the face position is roughly known from the landmark computation, we find a leaf node in the tree suited to represent the head position. Starting from this leaf, the hands and feet can be defined to be represented by those four leaves that are maximally far apart from the head and each other. Hence, once the head is determined, it is sufficient to identify four of those points to obtain candidates for hands and feet. Since the search of the feature points is based on the tree, the actual body pose during scanning does not change the result as long as no topological changes of the scan occur (e.g., a hand touches the body). Next, four additional feature points are equiped with semantic information (i.e., hand, feet, left and right). We separate the feet from the hands by choosing the two landmarks with the greatest tree-distance to the head as feet and the remaining two as hands. Left and right is decided by considering semantic vectors computed from the facial landmarks. Positional ambiguities introduced by crossing the arms are solved by tracing back the tree towards the shoulders. Similarly, a decision between left and right leg is made. In the last step, the skeleton tree is pruned by recursively removing all branches that do not start with one of the five feature points resulting in a clean skeleton (Fig. 3-(b), -(c)).

### 3.4 Post-proprocessing

In the post-processing step, the projection of a partially refined template onto the scan is performed. Two variants of the projection are possible: nearest-neighbour and reverse nearest-neighbour. During the nearest-neighbour projection, every point of the template is projected onto the nearest point of the scan.

[1] "semantic vector" is a vector giving meaning to a certain direction. The unit vector X does not have semantics but the vector "right" has a semantic, namely that it points to the right with respect to a reference.

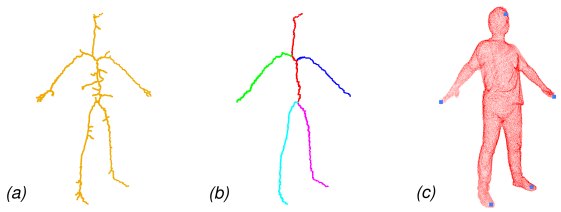


**Figure 3:** Extraction of the body landmarks. From left to right: (a) minimum spanning tree extracted from the scan using the Kruskals algorithm [18]; (b): a clean skeleton obtained by recursively removing all branches that do not start with one of the five feature points; (c): the resulting skeleton with an overlayed scan.

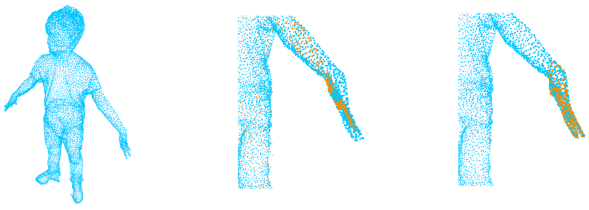


**Figure 4:** The figure shows a 3D scan (a), a result up to the step 3 (in orange) overlayed with the reconstruction (in cyan) (b) and the registration up to the step 3 (in orange) with registration correction (in cyan) (c).

This method often leads to point clustering. During the reverse nearest-neighbour projection, closest points on the template, for every point of the scan, are projected to the scan. This variant allows reducing point clustering effects and positively affects the overall registration quality. Fig. 5 illustrates differences in results depending on the respective projection algorithm.

### 3.5 Handling variety in poses

An accurate registration is particularly challenging in the hand region, due to the low point density compared to the amount of detail, and possible pose differences with the template. Moreover, missing data and different hand/finger configurations frequently occur. One example of a challenging case due to a bent hand can be observed in Fig. 4. In this case, non-rigid ECPD flattens the hand (see Fig. 4-(b)) and, consequently, the registration accuracy decreases.
A possible remedy is to use the hand of the rigidly registered template and perform non-rigid registration w.r.t. corresponding point cloud of the hand that was previously segmented via the framework. The result after applying the described technique can be observed in Fig. 4-(c).

It is noteworthy that the underlying ECPD algorithm is topology preserving. We exploit this property in multiple ways, i.e., in partial registration, in handling various topology, applying a topology transfer of the template, a straightforward transfer of vertex qualities like texture coordinates or skinning weights to the registered point cloud as well as the co-registration of auxiliary points (e.g., joint positions) allowing the transfer of rigging and skinning data between models. Furthermore, registration against the same 3D template guarantees a 1:1 mapping between all registration results, which is advantageous for keyframe based animations.

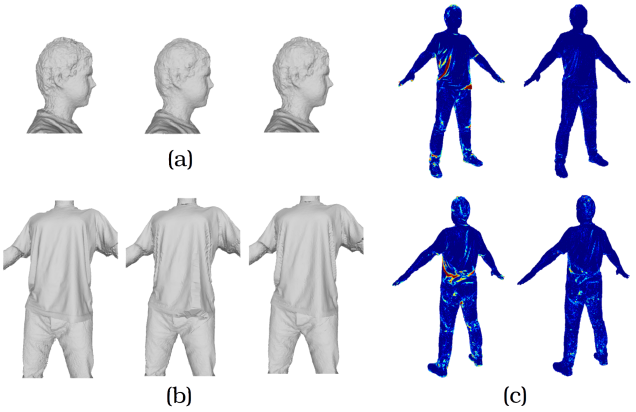


**Figure 5:** Accuracy evaluation of the proposed approach: (a) head of the original scan (left), registration result with the nearest-neighbour projection and result with the reverse nearest-neighbour projection (right); (b) torso of the original scan (left), registration result with the nearest-neighbour projection and result with the reverse nearest-neighbour projection (right); (c) detailed comparison of projection algorithms by the means of Hausdorff distance (Red>Yellow>Green>Blue scale): nearest-neighbour (left), reverse nearest-neighbour (right). Better viewed with zoom.

## 4 Experimental results

We run the proposed pipeline on a server under operating system Debian GNU 8.0 (codename Jessie), Intel Xeon E3-1245 V2 (3.4 GHz) processor and NVIDIA GeForce 660Ti graphics card (GK104-300-KD-A2 GPU).

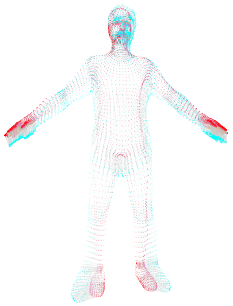

**Figure 6:** Point displacements from the initial template (cyan) to the final one (red).

**Implementation details**. Our semi-automated pipeline represents a `bash` script which invokes particular executables with automatically generated configuration files. Once a new scan needs to be processed, we invoke a script providing a path to the scan and the path to the data directory where results will be saved. The script manages all input data as well as configurations and feeds them to the respective executables. The executables accept config files which can guide all steps of the pipeline and can be changed for debugging, integration or performance evaluation purposes. Once the script is invoked, the pipeline starts with the rigid pre-alignment, followed by other steps until post-processing is accomplished.
The ECPD is implemented in C++/CUDA C and follows the description of [5, 6]. In the non-rigid case, we use the correspondence-preserving subsampling strategy proposed in [5] — instead of registering a reference with a scan directly, the template is subsampled and registered with the scan. Thereby, all prior matches are preserved and influence registration procedure. Further, the original template is registered with the subsampled template (the result of the previous step) and all points of the subsampled template are taken as prior matches. In this way, a linear speed-up is achieved. We adopt the subsampling strategy both for the global non-rigid and partial registrations. During the global registration, we also subsample the scan, since no decay in accuracy is observed, which results in even higher speed-up. In our supplementary material, we list the settings for particular steps of the proposed registration pipeline.

**Experiments on real data**. In total, we tested our approach on over hundred real-world scans with various appearance (clothes, hair, body metrics). Here, we show several results on the new data as well as on the FAUST data set [19]. The scans are obtained from a multiview system and reconstructed with an improved version of [20]. An exemplary scan with $4.66{\cdot}10^5$ points is shown in Fig. 1, on the left. All scans are captured in the pose similar to those of the human body template. While running, the template deforms as shown in Fig. 1, in the middle.

The result (Fig. 1, on the right) is very accurate. The appearance of the template is visually very close to the appearance of the original scan, despite that the template contains 2.65 times fewer points. Of course, not all details, especially in regions with a high total variation can be captured by the template. Fig. 5 provides a closer look to the result. As expected, even with the reverse nearest-neighbour projection, there are areas with a lower resolution of the structure. Fig. 6 shows point displacements from the initially rigidly registered template to the final non-rigidly aligned result. We make several observations about the displacement fields. First, there are no obvious distortions in the face area, despite the fact that the template has not matched the head after rigid pre-alignment. The leg area demonstrates a symmetric point drift, even though both parts are treated by the pipeline independently. The runtime for the processed scan from this experiment amounts to 2097 seconds. For other scans, the runtime lies in the interval {400; 2500} seconds depending on the number of points in the scans.

We also run the pipeline on several scans from the FAUST data set [19]. This data set serves for evaluation purposes of mesh registration algorithms with emphasis on varying poses, in inter- and intra-individual evaluation scenarios. Since we do not solve the problem of posing, we register several scans in similar poses to our segmented template. Results are shown in Fig. 7. The accuracy of results is in-line with the accuracy of other real-world scans. However, due to different hand poses (different from the pose of our template) and absence of prior matches, accuracy in the area of hands is lower.

## 5 Application of the proposed Framework in Treatment of Social Pathologies

The proposed pipeline can be adopted for generation of Animatable Virtual Avatars (AVA) resembling in appearance real persons. AVA are widely used in film industry and entertainment, augmented and virtual reality applications, and find its way nowadays into medicine and rehabilitation. Thus, our target application is an interactive system for curing pathologies such as schizophrenia or autism accompanied by social interaction burdens. The movement neuroscience and cognitive science suggest that it is easier for the respective patients to interact with subjects (real persons, virtual avatars or robots) looking similar to them. Thus, a system which allows generattion of AVAs and

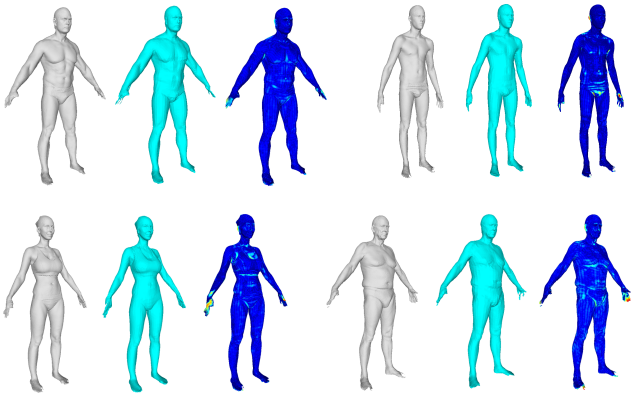

**Figure 7:** Results on the FAUST data set [19] shown in triples. For every triple: original scan (left), registration of the scan with the 10k template (middle), Hausdorff distance (Red>Yellow>Green>Blue scale) between the scan and our result. Our solution operates on 3D points, results shown as meshes for visualisation purposes.

changing their appearance gradually in a virtual environment is a concept for the next-generation therapy of this type of the social patalogies.

There are several key requirements on the target system. First, it should produce high-accuracy 3D scans. Second, to leverage real-time morphing (interpolation between multiple appearances) all scans must contain equal number of points. We experienced that the accuracy of modern affordable RGB-D sensors is too low to achieve the desired clinical effect, and opted for a multi-view reconstruction system consisting of multiple high-resolution RGB cameras. In contrast to modern RGB-D based systems which output a mesh from an implicit representation [21], multiview systems output point clouds. Thus, the subsequent design decision — adaptation of the proposed framework — is caused, on the one hand, by the necessity to accurately mesh the input scan. Due to the topology preserving the property of ECPD, the point topology of the template can be directly transferred to the registration result. On the other hand, the proposed pipeline gracefully solves two other tasks in one sweep — texturing of the AVA, co-registration of the joints for rigging (a core feature for animatibility of an AVA) and transfer of skinning weights. Since the original scan and the registered template are placed into the same coordinate system and are similar in appearance, the original high-resolution texture can be applied to the registered template. Moreover, the proposed framework is purely point-cloud based, and we are free to augment the template with the joints (a sparse point set).

Among over 140 generated AVA of the patients, only every 20-th AVA generation has required manual intervention (e.g., a fine adjustment of an elbow joint orientation or ECPD paremeters). All in all, our template-based solution and scanning in a pre-defined pose contributed to a well-balanced trade-off between the generality of the pipeline and high-quality results.

In Fig. 8, prototypes of the multi-view system and AVAs obtained with the proposed pipeline are shown. The AVAs are placed in a virtual environment mimicking the real surgery and can be moved interactively (e.g., by a pre-recorded motion sequence or a real-time capture of patient's movements). For further information on the application of the developed framework in medicine, please see links about the AlterEgo project in our supplementary material.

## 6 Conclusion

The methodology described in this paper can handle large point sets in reasonable time due to sophisticated sub-sampling strategies and a heterogeneous implementation. The proposed pipeline is robust and semi-automatic; it neither needs frequent user intervention nor parameter tuning. The pipeline can be applied to a broad range of applications and is to our knowledge the first method of its kind solely relying on point clouds. Most importantly, the possible quality of results is perfectly in line with the current state of the art competitor methods. Due to these benefits it opens up new application frontiers and effortlessly supports existent applications (e.g., [2]). Overall, the proposed pipeline provides a robust and automatic registration method able to produce satisfying results even in the case when no prior information can be provided. Adding prior information, if available, will additionally enhance the registration quality, but will not require any changes in the pipeline or the parameter settings. A current limitation of the pipeline lies in the requirement of small pose differences between the reference and template, as it is not pose-invariant. The introduction of partial registrations remedied the situation to a large extent but did not completely solve it. In future work, we are planning to combine the proposed framework with accurate kinematic motion capture [22]. Moreover, a special attention will be directed to the articulated pre-alignment to make the pipeline invariant of an initial body pose.

## References


[1] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM TOG **34** (2015) 248:1–248:16

[2] Wasenmüller, O., Peters, J.C., Golyanik, V., Stricker, D.: Precise and automatic anthropometric measurement extraction using template registration. In: 3DBST. (2015)

[3] Chen, Q., Koltun, V.: Robust nonrigid registration by convex optimization. In: ICCV. (2015)

[4] Ge, S., Fan, G.: Non-rigid articulated point set registration with local structure preservation. In: CVPR Workshops. (2015)

[5] Golyanik, V., Taetz, B., Reis, G., Stricker, D.: Extended coherent point drift algorithm with correspondence priors and optimal subsampling. In: WACV. (2016)

[6] Golyanik, V., Taetz, B., Stricker, D.: Joint pre-alignment and robust rigid point set registration. In: ICIP. (2016)

[7] Myronenko, A., Song, X.: Point-set registration: Coherent point drift. TPAMI (2010)

[8] Eckart, B., Kim, K., Troccoli, A., Kelly, A., Kautz, J.: Mlmd: Maximum likelihood mixture decoupling for fast and accurate point cloud registration. In: 3DV. (2015)

[9] Golyanik, V., Ali, S.A., Stricker, D.: Gravitational approach for point set registration. In: CVPR. (2016)

[10] Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. TPAMI **14** (1992) 239–256

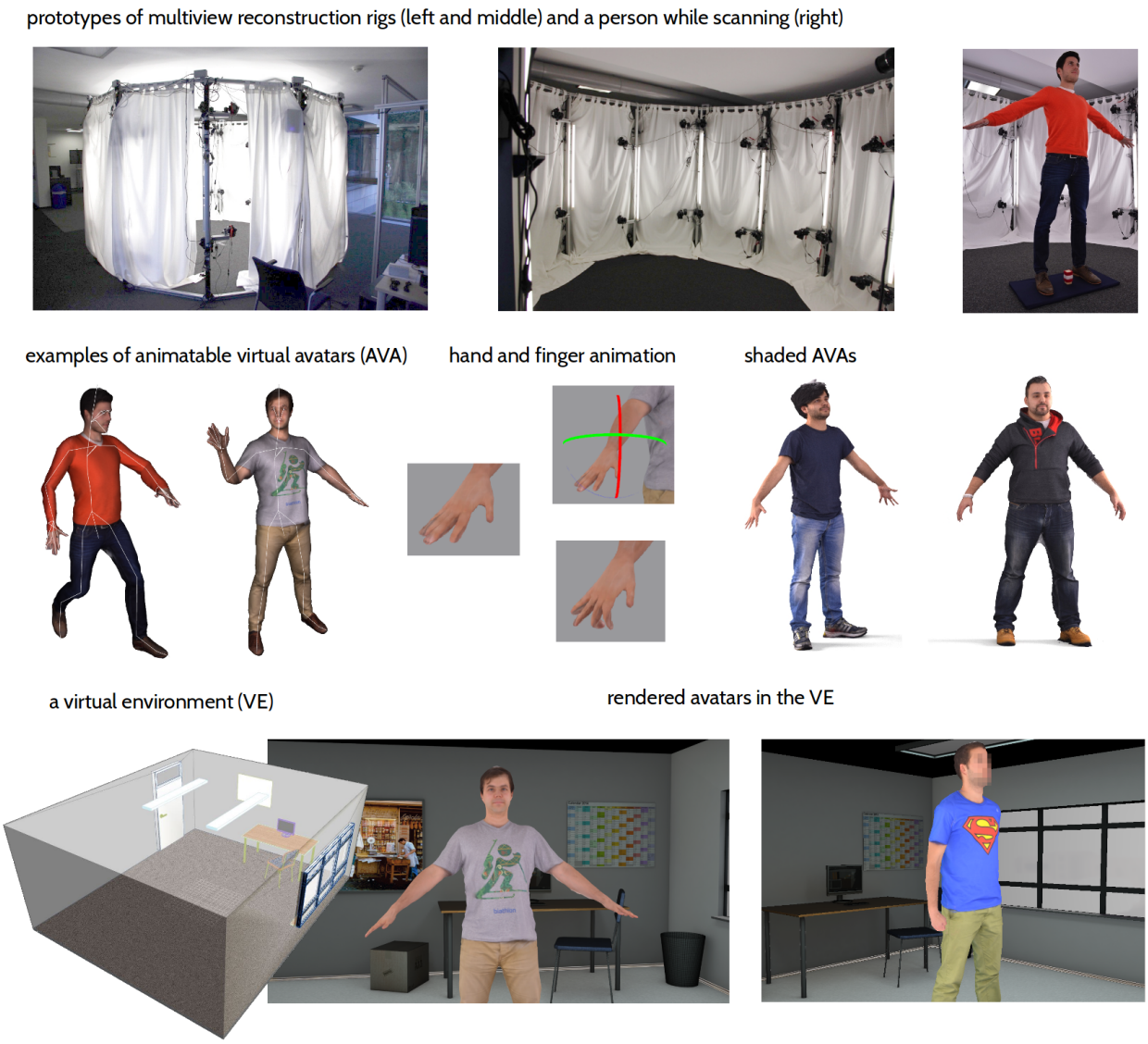


**Figure 8:** We adopt the proposed framework for registration of full 3D human body scans for the treatment of social pathologies. First, a patient is reconstructed on a high-resolution multiview RGB system. Next, a unified template with augmented joints is aligned with the body scan using the proposed framework. Next, the result is textured, and the skinning weights are computed. Finally, the AVA is optionally shaded and placed in a virtual environment. In the course of the therapy, the patient interacts with the AVA (e.g., repeats its movements) whose appearance changes gradually with time. The face of a real patient, bottom right, is pixelated.


[11] Pellegrini, S., Schindler, K., Nardi, D.: A generalisation of the icp algorithm for articulated bodies. In: BMVC. (2008)

[12] Ge, S., Fan, G., Ding, M.: Non-rigid point set registration with global-local topology preservation. In: CVPR Workshops. (2014)

[13] Dey, T.K., Fu, B., Wang, H., Wang, L.: Automatic posing of a meshed human model using point clouds. Computers & Graphics **46** (2015) 14 – 24

[14] Tsoli, A., Loper, M., Black, M.J.: Model-based anthropometry: Predicting measurements from 3d human scans in multiple poses. In: WACV. (2014) 83–90

[15] Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Incremental face alignment in the wild. In: CVPR. (2014) 1859–1866

[16] Palágyi, K., Kuba, A.: A 3d 6-subiteration thinning algorithm for extracting medial lines. Pattern Recognition Letters **19** (1998) 613 – 627

[17] Miklos, B., Giesen, J., Pauly, M.: Discrete scale axis representations for 3d geometry. ACM TOG **29** (2010) 101:1–101:10

[18] Kruskal, J.B.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. In: Proceedings of the American Mathematical Society, 7. (1956)

[19] Bogo, F., Romero, J., Loper, M., Black, M.J.: FAUST: Dataset and evaluation for 3D mesh registration. In: CVPR. (2014) 3794 –3801

[20] Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. TPAMI **32** (2010) 1362–1376

[21] Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: ISMAR. (2011) 127–136

[22] Taetz, B., Bleser, G., Miezal, M.: Towards self-calibrating inertial body motion capture. In: FUSION. (2016) 1751–1759