CLIPE 2024 - Creating Lively Interactive Populated Environments (2024)
N. Pelechano and J. Pettré (Editors)

Capture and Automatic Production of Digital Humans in Real Motion with a Temporal 3D Scanner

E. Parrilla[1], A. Ballester[1], J. Uriel[1], A.V. Ruescas-Nicolau[1] and S. Alemany[1]

[1]Instituto de Biomecánica de Valencia (IBV), Universitat Politècnica de València (UPV), Spain

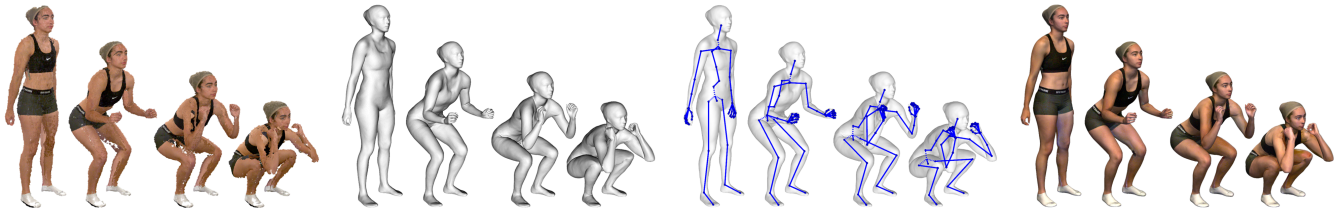


Figure 1: *Digital human production. From left to right: sequence of scanned point clouds, processed watertight meshes with vertex-to-vertex correspondences, rigged skeleton and color mapped textures.*

Abstract

The demand for virtual human characters in Extended Realities (XR) is growing across industries from entertainment to healthcare. Achieving natural behaviour in virtual environments requires digitizing real-world actions, a task typically laborious and requiring specialized expertise. This paper presents an advanced approach for digitizing humans in motion, streamlining the process from capture to virtual character creation. By integrating the proposed hardware, algorithms, and data models, this approach automates the creation of high-resolution assets, reducing manual intervention and software dependencies. The resulting sequences of rigged and textured meshes ensure lifelike virtual characters with detailed facial expressions and hand gestures, surpassing the capabilities of static 3D scans animated via separate motion captures. Robust pose-dependent shape corrections and temporal consistency algorithms guarantee smooth, artifact-free body surfaces in motion, while the export capability in standard formats enhances interoperability and further character development possibilities. Additionally, this method facilitates the efficient creation of large datasets for learning human models, thus representing a significant advancement in XR technologies and digital content creation across industries.

CCS Concepts

• ***Computing methodologies*** → ***Shape modeling;*** *Machine learning; Computer vision;* ***Animation;*** *Image manipulation;* ***Graphics file formats; Mesh models; Mesh geometry models;*** *Image processing;* ***Computer vision problems; Reconstruction;***

1. Introduction

Advances in technology are making possible that Extended-Realities (XR) are becoming more present in our daily lives at different levels. Leisure, media and entertainment industries rely heavily on digital artists and creators, and on virtual-, augmented- and mixed-realities (VR/AR/MR). Every day, there are more products and environments that are not just designed virtually using CAD software, but that can also be simulated with digital humans in VR or can be interacted in AR [Mor98; Ghe98; LJLS19]. Health, culture and education are also increasingly being offered in immersive and virtual experiences [AJGA23; WFD*22; BMR23; CPWD21]. The creation of digital content, and in particular the creation of virtual characters, is therefore of central importance. Modelling and bringing them to life requires highly specialized programming combined with artistic skills and considerable time and cost investments.

Making virtual human characters look, move, behave and interact more naturally is subject to continuous focus [MT05; DMT12]. One way of making it possible is to capture the appearance, motion, behaviour and interaction of actual people and then to confer

DOI: 10.2312/cl.20241048

it to virtual characters. The current practice and software used to create virtual human characters varies depending on the field of application but in essence, it consists of either capturing one or several static body shapes using 3D scanning technologies [DvdW98; DT13] and animating them using motion and expression data captured separately [BP07; DTSS07]; or capturing body surfaces in motion using volumetric video or 4D scanning systems. However, the capture of such data and its curation to make them usable in production requires intensive work by digital artists and the combination of multiple hardware and software. In the past years, the modelling of human shape, motion and expression to accurately represent the full body detail at the level of shape, pose, facial expression, and hand manipulation has been subject of important research and development efforts [ACP03; HLRB12; LMR*15; XBZ*20]. Yet, the curation of the captured data for learning new human models with such "natural" capabilities requires very specialized skills in Computer Vision (CV), Computer Graphics (CG) and Machine Learning (ML) which are beyond the typical skills of digital artists, designers, engineers, ergonomists, and other specialists in the fields of entertainment, product design, human factors, culture or education. In either case, the curation steps include the registration of the 3D point cloud (or sequences of point clouds) to a specific mesh topology that is afterwards rigged and that has to be able to be textured, animated and rendered.

This paper describes a new technology, Move4D [PBP*19], that tackles the main barriers and inefficiencies in the process of digitizing a real human in motion to deliver a high resolution 3D virtual character ready to be used in the XR applications in different fields such as virtual garment development and fit simulation, virtual ergonomics, media production, video-games, simulation of sports and clinical biomechanics, or the learning of human body models among others. The proposed technology, is an end-to-end solution from the sensors that made up the studio or laboratory able to capture the 3D body surface in motion, to the export of the data in file exchange formats. To make it possible, it includes a fully automatic registration algorithm to process the acquired raw data. It produces a sequence of rigged and textured meshes, herein homologous meshes, with vertex-to-vertex correspondence between subjects and along the movement sequence (see Figure 1). The term 'homologous mesh' was firstly introduced by [MK00] and consists of representing human bodies (or body parts) with meshes with the same number of vertices and topology, where each data point is defined based on the anatomical homology of humans.

2. Background

3D temporal capture for XR applications. Different XR applications may have different requirements for the capture of human body surfaces in action. Generally, the capturing solution should be able to operate at high frequencies (i.e. 60 fps or more) and instantly (i.e. 10 ms or less) to get fast actions, at high resolution (i.e. at least in the order of millimetres) to capture expressions and accurate shapes and features, and it should be able to cover a large scanning volume (2 m^3 or more) to let the person move freely and perform the action comfortably. Capturing body surface complying with these requirements is challenging, especially, if such capture is not limited to a point cloud and should also include other information such as high resolution images to create textures and support the registration process using Deep Learning (DL) models. Furthermore, being able to optimize the balance between requirements depending on the application would also be practical and desirable, for instance, increasing the frequency for the sake of resolution or vice-versa.

Creation of virtual characters. The creation of virtual characters for XR applications from 3D scans requires using different 3D modelling software that assists technicians and digital artists in a series of complex and specialized manual operations. The baseline operations to curate and integrate the captured data (i.e. 3D points or surfaces, images and motion capture data) are: (i) obtaining a watertight mesh (or meshes, in the case of temporal 3D scanning or capturing several static poses) with the adequate topology (i.e. number of elements, distribution and type of elements), (ii) creating a UV map and using the captured images to create a texture, (iii) skinning, which is equipping the mesh with a skeleton (i.e. joint hierarchy, joint 3D positions, joint axes orientation, and weight painting), (iv) in case the motion has been captured separately, it has to be adopted and adapted to the skeleton of the character, (v) in case there are multiple meshes including actual surface deformations that are required to be preserved during skeletal animation, they have to be retrieved as corrective shapes from the corresponding skeletal animation or pose. Depending on the application, the characters are retouched on any of the previous features (i)-(v) or enriched with virtual garments, hair, and eventually with advanced texture and sub-texture maps. Some of the most popular general purpose 3D modelling software are Maya [Autc], Blender [Ble] or 3ds Max [Auta], each of them offering different particular advantages to the users. Other popular specialised software examples are Marvelous Designer [CLO] for garments, Substance 3D Painter [Ado] for advanced textures or Faceform [Fac], 3DCoat [Pil] or TopoGun [Pixb] for mesh topology. Finally, the characters should be imported into a computing platform or game engine to make part of the production project. The most common platforms are Unreal Engine [Epi], Unity [Uni] and Omniverse [NVI].

Relevant file exchange formats. Within this context, the use of file exchange formats is essential to exchange data from one software to another and to pack the final result and integrate it to the project. There are several data formats that are able to represent complex 3D animations using meshes with texture encompassing skeletal animation with vertex animation. Among these formats we have FBX [Autb], glTF [Khrb] and USD [Pixa]. FBX is a closed binary that requires to use the SDK provided by Autodesk to use it, while USD and glTF are open formats that can be written in ASCII or binary. glTF also has an official tool for the verification of the compatibility of the files [Khra].

3. Capture of actual human surfaces in real motion

The capture system we have used is a temporal 3D scanner composed of a set of calibrated sensors placed around the scene to capture the different points of view. This system is modular and highly configurable, allowing to adjust the number of sensors and their location according to the needs of the scene and the motion to be acquired. The sensor is made up of a pair of infrared (IR)

cameras and an IR projector for 3D acquisition, a RGB camera for color capture and an independent processing unit (CPU). The cameras have a resolution of 2048x2048 pixels (1024x1024 pixels at medium resolution) and a maximum frame rate of 90 fps (178 fps at medium resolution). The IR projector has the same lens and angle of view/projection as the cameras, allowing to easily adjust the angle of capture by changing the focal length of the lenses. In addition, the distance between the IR cameras, their orientation and exposure time are configurable, making it easy to adapt the sensors to the desired acquisition volume and resolution. Once the images have been captured, the sensor's CPU processes the IR images to obtain the depth map and the point cloud of the scene at a speed of 15 fps. For each frame, along with the RGB image and calibration information, the point cloud is sent to the scanner control server as soon as it is available. The server joins the partial data from the different sensors to form the full sequence. The system also integrates a visible lighting system to optimize the color acquisition. The scanner can also be synchronized with other laboratory equipment such as force platforms, accelerometry, pressure mats, photoelectric sensors, etc. as well as with other vision systems such as MoCap, RGB or thermal cameras.

We have used a 16-sensor configuration that allows the capture of a 2x3x3m volume with a spatial resolution of ∼1mm and an exposure time of 1ms. Using more sensors may be desirable for certain applications or movements to reduce the number of occlusions and thus increase the captured surface area. The raw data captured with this configuration produces a dense point cloud of more than 4 million points per frame.

4. Processing of temporal sequences of scans

This processing converts the 3D raw data into a noise- and artefact-free watertight dense mesh with texture. This homologous mesh is rigged and has vertex-to-vertex correspondence between subjects and along the movement sequence. This processing is based on the registration of a human articulated model to the 3D point cloud captured by the scanner, using a common 3D template for all scans.

4.1. Human Models

Human shape and pose models are fundamental tools in various applications, ranging from entertainment to healthcare. Shape models based on Principal Component Analysis (PCA) [ACP03] focus on capturing anatomical shape under static conditions, employing geometric modeling techniques to represent human body morphology. On the other hand, articulated shape and pose models [ASK*05; HLRB12; LMR*15; XBZ*20] incorporate the dynamic aspects of human motion, utilizing articulated skeletons to depict body postures and gestures, making them essential in animation, virtual reality, and biomechanical research.

Body model. We represent the human model, $M(\beta,\ \theta)$, as an articulated mesh composed by 49530 vertices, specified by a skeleton with 63 joints and skin deformed based on standard linear blending skinning (LBS) function, $W\left(T_B\ (\beta,\theta),\ J\left(T\left(\beta\right)\right),\theta,\Omega\right)$, as is shown in Figure 2.

$$M(\beta,\theta) = W\left(T_B\ (\beta,\theta),\ J\left(T\left(\beta\right)\right),\theta,\Omega\right) \tag{1}$$

$$T_B(\beta,\theta) = T(\beta) + B(\beta,\theta) \tag{2}$$

$$T(\beta) = \bar{T} + S\beta \tag{3}$$

where β and θ are the shape and pose (Rodrigues angles) parameter vectors, respectively, $\bar{T}$ and S are the mean and the loading matrix of a PCA of a dataset of body shapes in rest pose, $B(\beta,\theta)$ is a function that models pose-dependent soft tissue deformations, $J(T(\beta))$ is a regression function that computes joint locations, and Ω are the skin attachments of the articulated model. The pose parameter vector $\theta = \{\theta_B, \theta_{HR}, \theta_{HL}\}$ contains the parameters for the body, θ_B, and for the right and left hands, θ_{HR} and θ_{HL}.

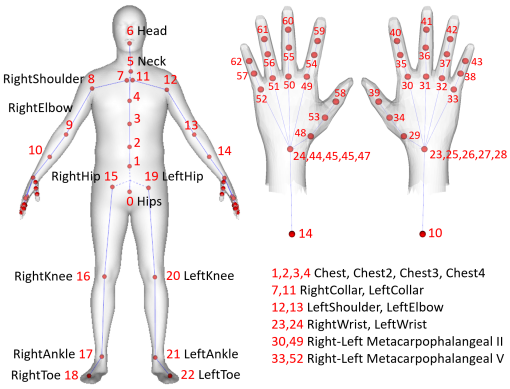


Figure 2: *Template mesh and 63-joint rigged skeleton of the body model. Names of the main joints are shown.*

All datasets used in the creation of the body model have been generated using simplified versions of the method described in section 4.2. This process is iterative, progressively integrating improvements as they become available.

The shape model $T(\beta)$ has been computed from the CAESAR dataset [RDP99] and a Spanish population dataset [BVN*15], totalling 3565 registrations for males and 8731 for females. Each dataset contains meshes with the same topology as our template that have been aligned to high-resolution 3D scans using a simple human model consisting of an articulated template and manually placed body landmarks.

The soft tissue model $B(\beta,\theta)$, skin attachments Ω, and joint regression matrix $J(T(\beta))$ have been trained using a dataset of 24213 poses from 11 subjects captured with the temporal scanner. This dataset has been processed employing an initial joint regression matrix computed from a skeleton based on International Society of Biomechanics (ISB) criteria [WC95; WSA*02; WvD*05], artist-generated skin attachments, and assuming zero soft tissue deformation.

To improve the robustness of the registration, a set of parameters describing in a simplified way the predominant body poses has been computed. These parameters have been obtained from a PCA of the pose parameters θ extracted from a dataset consisting of 196646 scans of 113 subjects performing different movements, processed using a model incorporating the previously trained soft

tissue model $B(\beta,\theta)$, skin attachments Ω, and joint regression matrix $J(T(\beta))$.

$$\theta_B(\varphi_B) = \bar{\theta}_B + P_B\varphi_B \tag{4}$$

where φ_B is the body angle parameter vector, and $\bar{\theta}_B$ and P_B are the mean and loading matrix of a PCA of the body pose dataset.

Hand model. Analogous to the body model, an articulated right hand model has been computed using 2018 hands from the MANO dataset [RTB17]. The template used is a subset of the body template formed by 1315 vertices of the hand region. We subsequently create a left hand model by mirroring the computed right hand model. These models have been merged with the body model, incorporating the hands joint regression matrix into the full body regression matrix, and integrating the skin attachments and soft tissue model using smoothness conditions in the joint regions of body and hand models.

As in the body, the hand poses have been parameterized using a PCA to reduce the dimensionality of the data, which allows for a more concise representation of the intricate variations in hand movements, facilitating the synthesis of hand pose.

$$\theta_{HR}(\varphi_{HR}) = \bar{\theta}_{HR} + P_{HR}\varphi_{HR} \tag{5}$$

where φ_{HR} is the right hand angle parameter vector, and $\bar{\theta}_{HR}$ and P_{HR} are the mean and the loading matrix of a PCA of the hand pose datasets.

Afterward, we mirror the computed right hand pose model to generate its corresponding left hand model.

$$\theta_{HL}(\varphi_{HL}) = \bar{\theta}_{HL} + P_{HL}\varphi_{HL} \tag{6}$$

Face model. The Basel Face Model (BFM) [GMB*18] is a highly detailed parametric three-dimensional representation of the human face. This model addresses both structural aspects and facial expressions, making it versatile for various applications such as realistic image synthesis and facial analysis. The BFM can be expressed through mathematical equations describing facial geometry. A simplified form of the model uses a linear combination of shape and expression vectors, represented as:

$$F(\alpha_S,\alpha_E) = \bar{F} + F_S\alpha_S + F_E\alpha_E \tag{7}$$

where $\bar{F}$ is the mean shape, F_S and F_E are shape and expression matrices respectively, and α_S, α_E are coefficients determining facial deformations. This parametric approach allows for a compact and efficient representation of facial variability.

In the registration process, we use a modification of the BFM model to adapt and align it to a dense set of vertices of the face of our template, thus making it compatible with our human model.

4.2. Automatic registration

Registration consists of calculating the shape and pose of the subject in each of the frames that compose the 4D sequence. To make the computation more efficient, we use a short capture of the subject in a pose close to the rest pose, which we call A-Pose, which is acquired in addition to the motion sequences. In this way, the shape parameters are computed from this A-Pose and are fixed for the remaining motion sequences, simplifying the computation since it is only necessary to compute the pose parameters in every frame.

To start the registration process, it is necessary to calculate a set of surface landmarks and joint landmarks to guide the process. In the case of the body, for the A-Pose processing, we use a PointNet++ [QYSG17] based network, trained with 22826 scans of 11 subjects performing different movements and the 12296 static scans from the CAESAR and the Spanish datasets. This network, that is applied directly to the raw data, classifies the vertices of the point cloud associating each of them to a vertex of our template. The network has been trained using a pointwise softmax cross-entropy loss function. The dataset used has been generated by associating each vertex of the point cloud with the nearest vertex of its corresponding homologous mesh. These meshes have been computed using the method described in section 4.2, using a sparse set of manual landmarks for processing the A-Poses. This PointNet++ based network is very robust in poses close to the rest pose and also provides information about the subject's body shape. However, it is not robust in very complex poses or when large body regions are missing due to occlusions.

For the motion sequences, we compute a set of joint landmarks employing standard pose estimation networks [BGR*20], that have been shown to be very robust but do not provide information about the body shape. These networks are applied to a set of virtual images rendered from the colour point cloud obtained with the scanner corresponding to different camera observation viewpoints. Images are generated by rotating the virtual camera around the scan at three different heights, thus capturing different views of the subject. We then run automatic detectors to identify 2D landmarks in rendered images. Once such estimates are available, we can triangulate to obtain the 3D landmarks.

In the case of hands and face, for both A-Pose and sequences, we use the landmarks detectors [ZBV*20; GAK*20] directly on the images acquired by the RGB cameras of the scanner sensors.

The registration is controlled by the model, but we introduce some optimization additional terms that allow us to fit any human shape even if it is not defined by the shape and pose model. Thus, we modify equations 1, 2 and 3 as follows:

$$\begin{aligned} M(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f) = \\ W(T_B(s,\beta,\theta,\varepsilon_s), J(T(s,\beta,\varepsilon_s),s_b),\theta,\Omega) + \varepsilon_f \end{aligned} \tag{8}$$

$$T_B(s,\beta,\theta,\varepsilon_s) = T(s,\beta,\varepsilon_s) + sB(\beta,\theta) \tag{9}$$

$$T(s,\beta,\varepsilon_s) = s(\bar{T} + S\beta) + \varepsilon_s \tag{10}$$

where s is a scaling factor applied to the shape of the subject, s_b are scales applied to each bone of the skeleton and ε_s and ε_f represent changes in shape and pose, respectively, that the model cannot explain for a specific subject and frame. ε_s is the same for all the frames of the same subject, while ε_f is specific for each frame.

The s_b parameters are only optimized in the A-Pose initial alignment phase. The introduction of these parameters allows to modify the length of the different body limbs of the template so that it is optimally aligned with the target pose. In the rest of the process they are fixed and equal to 1.

The registration consists of finding the parameters s, β, θ, ε_s and ε_f that generate a mesh equivalent to the captured 3D point cloud.

Initial alignment. Alignment consists of finding initial Rodrigues angles θ and scales s_b and s that fit the model closer to the scan. In the A-Pose capture, the parameters β and ε_s are set to a value of zero. In all other sequences, the s, β and ε_s parameters are set to the values obtained in the A-Pose processing of the captured subject.

In the alignment, 3D landmarks of hands L_{HR} and L_{HL}, face L_F and body L_B are used. Note that L_B differs between A-Pose and sequences, as previously mentioned. Initially, a scaled rigid alignment of a set of landmarks is performed to compute the value of s, which is fixed during the rest of the processing.

The alignment is divided into three phases. In the first phase, an alignment is performed in which just the parameters φ_B, φ_{HR} and φ_{HL} of the body and hand pose PCAs are directly optimized but not the angles. This allows us to make an initial approximation to the target pose by constraining the body motion to natural movements.

To achieve all this, we minimize an objective function consisting of a landmark term, E_{Ln}, and a temporal consistency regularization term E_T defined below. The landmark term penalizes the squared Euclidean distance between surface and joint landmarks and the corresponding model vertices and skeleton joints.

$$\begin{aligned} E_{L1}\left(s,\beta,s_b,\varepsilon_s,\varphi_B,\varphi_{HR},\varphi_{HL},\varepsilon_f,T_g,L_B,L_{FH}\right) = \\ \lambda_B \sum_{i=1}^{N_B} \left\| M_{Bi}\left(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f\right) - L_{Bi} + T_g \right\|^2 + \\ \lambda_{FH} \sum_{i=1}^{N_{FH}} \left\| M_{FHi}\left(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f\right) - L_{FHi} + T_g \right\|^2 \end{aligned} \tag{11}$$

with $\theta = \left\{\overline{\theta}_B + P_B\varphi_B, \overline{\theta}_{HR} + P_{HR}\varphi_{HR}, \overline{\theta}_{HL} + P_{HL}\varphi_{HL}\right\}$, where $L_{FH} = \{L_F, L_{HR}, L_{HL}\}$, N_B is the number of body landmarks and joints, N_{FH} is the number of face and right and left hands landmarks and joints and T_g is the global translation. $M_B(\cdot)$ and $M_{FH}(\cdot)$ refer to the vertices/joints of the model $M(\cdot)$ corresponding to the landmarks L_B and L_{FH}, respectively.

In the second phase, the body angles are directly optimized along with the parameters of the hand pose PCAs.

$$\begin{aligned} E_{L2}\left(s,\beta,s_b,\varepsilon_s,\theta_B,\varphi_{HR},\varphi_{HL},\varepsilon_f,T_g,L_B,L_{FH}\right) = \\ \lambda_B \sum_{i=1}^{N_B} \left\| M_{Bi}\left(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f\right) - L_{Bi} + T_g \right\|^2 + \\ \lambda_{FH} \sum_{i=1}^{N_{FH}} \left\| M_{FHi}\left(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f\right) - L_{FHi} + T_g \right\|^2 \end{aligned} \tag{12}$$

with $\theta = \left\{\theta_B, \overline{\theta}_{HR} + P_{HR}\varphi_{HR}, \overline{\theta}_{HL} + P_{HL}\varphi_{HL}\right\}$.

In the third phase, the body landmark term is replaced by the distance between a sparse set of scan vertices and the nearest model vertices. This allows correcting the differences between the position of the body landmarks and the actual position of the skeleton joints, while improving the pose in regions where landmarks are scarce or non-existent (e.g., spine).

$$\begin{aligned} E_{L3}\left(s,\beta,s_b,\varepsilon_s,\theta_B,\varphi_{HR},\varphi_{HL},\varepsilon_f,T_g,Y,L_{FH}\right) = \\ \lambda_B \sum_{i=1}^{N_Y} \min \left\| M\left(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f\right) - Y_{Bi} + T_g \right\|^2 + \\ \lambda_{FH} \sum_{i=1}^{N_{FH}} \left\| M_{FHi}\left(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f\right) - L_{FHi} + T_g \right\|^2 \end{aligned} \tag{13}$$

where Y is the target scan and Y_B is a sparse set of vertices.

During alignment, physiological constraints are used in the optimization of the angles according to the possible movement of each joint (one, two or three degrees of freedom). On the other hand, a temporal consistency term is also used between frames that penalizes the differences between the parameters to be optimized. The function of this term is to reduce the jitter between consecutive frames and to resolve potential posture ambiguities that may occur in some frames. For example, if the elbow is extended it is difficult to determine whether the rotation of the hand has occurred in the shoulder or in the forearm.

$$E_T(p) = \sum_{i=1}^{N_p} \sum_{j=2}^{N_f} f_s \left\| p_i^j - p_i^{j-1} \right\|^2 \tag{14}$$

where p is the total set of parameters to be optimized in each phase, N_p is the total number of parameters, N_f is the number of frames and f_s is the acquisition frequency of the scanned sequence, so the weight of this term is greater as the capture frequency increases. p_i^j refers to the i-th parameter to be optimized for frame j.

If we denote E_{Ln}^j to the marker distance errors in frame j, the total energy for the initial alignment is as follows:

$$E_A(p) = \sum_{j=1}^{N_f} E_{Ln}^j + \lambda_T E_T \tag{15}$$

Finally, we can compute the first approximation to the scan A for each frame as:

$$A = M\left(s,\beta,s_b,\varepsilon_s,\theta,\varepsilon_f\right) + T_g \tag{16}$$

Optimization. The optimization consists of finding the parameters β, θ, ε_s and ε_f that define the target scan. In the case of A-Pose, all parameters are optimized. In the case of the sequences, β and ε_s are fixed at the values computed for the A-Pose of the subject and only the pose parameters θ and ε_f are optimized. For simplicity and clarity, in this section we will refer to the model in equation 8 as $M(\beta,\varepsilon_s,\theta,\varepsilon_f)$, since the parameters s_b are set to 1 and remain constant, and s also remains constant at the value computed in the initial alignment.

During this process ε_f encompasses both shape changes and global translation of the scan. The optimization is an iterative process, in which the error functions E_D, E_{ES} and E_{EF} described bellow are minimized by modifying the weights of each one at each iteration k. The solution of each iteration involves solving a minimization problem using iterative nonlinear optimization algorithms.

We start by initializing the angles θ to those computed in the

initial alignment process and ε_f as:

$$\varepsilon_f = A - W\left(T_B\left(s,\beta,\theta,\varepsilon_s\right),\ J\left(T\left(s,\beta,\varepsilon_s\right)\right),\theta,\Omega\right) \tag{17}$$

In the remaining iterations k, all the vertices of ε_f are initialized to the translation T_g, resetting the effect of ε_f on the change of shape. This allows correcting the errors that may occur in the associations of the vertices of the model with the vertices of the point cloud, thus avoiding the propagation of mismatches that may occur between the model and the scan in the first iterations of the processing.

$$E_D(\beta,\varepsilon_s,\theta^{k-1},\varepsilon_f,Y) = \min\left\|M(\beta,\varepsilon_s,\theta^{k-1},\varepsilon_f) - Y\right\|^2 \tag{18}$$

where θ^{k-1} is the angle parameter vector computed in the previous iteration $k-1$. E_D penalizes the squared Euclidean distance between scan vertices and model vertices.

$$E_{ES}(\varepsilon_s) = \sum_e \left\|\varepsilon_{si} - \varepsilon_{sj}\right\|^2 \tag{19}$$

$$E_{EF}(\beta,\varepsilon_s,\theta^{k-1},\theta,\varepsilon_f) = \sum_e \left\|W_e(s,\beta,\theta^{k-1},\varepsilon_s) + (\varepsilon_{fi} - \varepsilon_{fj}) - W_e(s,\beta,\theta,\varepsilon_s)\right\|^2 \tag{20}$$

$$\begin{aligned} W_e(s,\beta,\theta,\varepsilon_s) = {} & W_i\left(T_B(s,\beta,\theta,\varepsilon_s), J\left(T\left(s,\beta,\varepsilon_s\right)\right),\theta,\Omega\right) \\ & - W_j\left(T_B(s,\beta,\theta,\varepsilon_s), J\left(T\left(s,\beta,\varepsilon_s\right)\right),\theta,\Omega\right) \end{aligned} \tag{21}$$

where i and j denote the two vertices of an edge e of the model. E_{ES} enforces smoothness in the shape term of the model. The error function E_{EF}, in addition to providing a smoothness condition controlled by the model, ensures that the new angles θ try to explain most of the pose-dependent shape variation, minimizing the contribution of ε_f to the shape change in each iteration. Note that since it is a function that operates on the edges, the global translation contained in ε_f has no effect.

Due to the unique characteristics of the face, which may undergo changes in expression and gestures during capture, it is critical to process it properly to maintain an accurate correspondence between vertices throughout the sequence. In addition to the previously mentioned error terms, other specific error functions E_L and E_{FM} are incorporated to address this additional complexity.

$$E_L(\beta,\varepsilon_s,\theta^{k-1},\varepsilon_f,L_F) = \sum_{i=1}^{N_F}\left\|M_{Fi}(\beta,\varepsilon_s,\theta^{k-1},\varepsilon_f) - L_{Fi}\right\|^2 \tag{22}$$

where $M_F(\cdot)$ refers to the vertices of the model $M(\cdot)$ corresponding to the landmarks of face L_F and N_F is the number of landmarks. E_L penalizes the squared Euclidean distance between landmarks of face and the corresponding model vertices, as in the initial alignment.

$$\begin{aligned} & E_{FM}\left(\beta,\varepsilon_s,\theta,\varepsilon_f,\alpha_S,\alpha_E,T_f\right) = \\ & \left\|M_{FM}\left(\beta,\varepsilon_s,\theta,\varepsilon_f\right) - \left(R_{head}F\left(\alpha_S,\alpha_E\right) + T_f\right)\right\|^2 \end{aligned} \tag{23}$$

where $M_{FM}(\cdot)$ are the dense set of face vertices corresponding to the face model $F(\alpha_S,\alpha_E)$ and R_{head} is the rotation given by the head joint. E_{FM} controls that the distribution of the face vertices is consistent with the face model. In the A-Pose, the translation of the face T_f, α_S and α_E are optimized, while in the sequences the shape parameter vector α_S is fixed at the values computed in A-Pose, allowing only changes in the expression parameters α_E.

Finally, the total energy for the optimization is as follows:

$$\begin{aligned} & E_O\left(\beta,\varepsilon_s,\theta,\varepsilon_f,\alpha_S,\alpha_E,T_f,L_F,Y\right) = \\ & \lambda_D E_D + \lambda_{ES}E_{ES} + \lambda_{EF}E_{EF} + \lambda_L E_L + \lambda_{FM}E_{FM} \end{aligned} \tag{24}$$

Thus, the registration of the scan is given by $M(\beta,\varepsilon_s,\theta,\varepsilon_f)$. We can also compute the rest pose of the subject with equation 10 and generate the skeleton for each frame using the optimized angles θ and the global translation $T_g = \bar{\varepsilon_f}$ computed as the mean of ε_f.

Texture. Our template has a UV map for the projection of texture images on the 3D object. This UV map is common for all the scanner registers due to the vertex-to-vertex correspondence and the common topology of all the processed captures. From the RGB images and calibration data, including intrinsic and extrinsic camera parameters, a projection of the RGB images onto the previously defined UV map is performed using standard texture mapping techniques [WMG14]. This process allows the visual characteristics of the RGB images to be accurately mapped to the UV coordinates of the processing, thus generating a coherent and realistic texture that perfectly matches the geometry of the object in question. In addition, the use of a common UV map for all models ensures visual consistency between the generated textures.

5. Use of the registered data to create virtual characters

The sequence of registered meshes resulting from the algorithms described in section 4 can be packed into standard containers because the meshes have vertex-to-vertex correspondence along the sequence, they share a common skeletal parameterization (i.e. joint hierarchy, skin attachments and rest pose) and UV map. Before exporting sequences in FBX and glTF formats, it is necessary to convert the pose deformations to fit the format supported by these containers. Both formats store the mesh in its rest pose computed from $T(s,\beta,\varepsilon_s)$, along with the skeleton joints computed from $J(T(s,\beta,\varepsilon_s))$ and the corresponding skin attachments Ω. Additionally, for each frame of the sequence, the translation T_g and the angle values θ that define the mesh pose are stored, alongside a term added to the rest pose mesh, which accounts for shape changes due to the pose. This term, that we call ε_r, can be calculated for each frame as follows:

$$\begin{aligned} \varepsilon_r = {} & W^{-1}\left(M(s,\beta,\varepsilon_s,\theta,\varepsilon_f) - T_g, J\left(T\left(s,\beta,\varepsilon_s\right)\right),\theta,\Omega\right) \\ & - T(s,\beta,\varepsilon_s) \end{aligned} \tag{25}$$

where $W^{-1}(\cdot)$ is the inverse of the linear blend skinning function $W(\cdot)$.

The nomenclature of the corrective shapes ε_r in FBX is *blendshape* and in glTF it is *Morph Targets*. Blendshapes in FBX can be made up of one or several shape targets thus offering different ways of animating ε_r at each frame. One way, herein *multi-blendshape* variant, is to create one blendshape with a single target per frame and then define the animation as a sequence of activations of one frame and deactivation of the rest. Another way, herein *single-blendshape* variant, is to use a single blendshape with multiple targets and then define the animation as the sequence of targets.

Morph Targets in glTF contain one shape target per frame, as in FBX multi-blendshape variant.

Computation of metrics. In addition to the virtual characters, some VR applications such as virtual product design and ergonomic simulations, which methods derive from the use of traditional anthropometry, still require or can benefit from the computation of different body metrics in their processes. The fact of having homologous meshes with vertex-to-vertex correspondence between subjects and with a common skeletal parameterisation facilitates considerably the implementation of any kind of body metric that can be computed from the body surface and/or the skeletal structure [BPU*14; YZF*24]. One example is the virtualization of a measuring tape following ISO or ASTM guidelines [Int17b; Int17a; AST15] and using a subset of vertices learned as the anatomical references [RDBJ24]. Another example is the computation of surfaces, volumes and inertial moments for the whole body or for body parts that can be obtained by groups of mesh elements or by lines defined by edges or vertex sequences. Moreover, the fact that such anatomical correspondence is preserved along the motion sequence it makes it possible not only to generalize such measurements to the dynamic field [URI*22] but also to compute classic biomechanical metrics such as joint angles, moments and forces [DVP*22; PDL*23; RDB*22].

6. Results and discussion

The described algorithms were implemented as a C++ library which allows easy integration into any software. In addition, the use of this language allows optimal memory management and high computational efficiency making possible the processing of long sequences that require a large amount of resources. Total processing time is about 13 seconds per frame on an Intel Core i9-13900 CPU (2 GHz/5.6 GHz, 32 threads), including point cloud sequence capture and computation of the homologous meshes, skeletons and textures.

6.1. Evaluation of the registration process

A comprehensive experimental study was carried out to evaluate multiple aspects of the algorithm, including its accuracy, robustness and convergence capability. In addition to analyzing the numerical results of the processing error, special attention was paid to the visual evaluation of the resulting meshes. This visual inspection was fundamental to evaluate the level of realism achieved and to examine thoroughly the perceptual details. In addition, the importance of assessing the temporal coherence between the different frames of the sequence was emphasized, which plays a crucial role in the overall quality of the results.

Quantitative analysis. To evaluate the effectiveness of the registration process, a quantitative analysis was performed that involved calculating the distances between the vertices of the registered mesh and the vertices of the point cloud in a motion sequence covering the entire scanning volume, as illustrated in Figure 3. The results showed a computed average error of $\sim$1 mm, matching the spatial resolution of the scanner configuration used. This agreement highlights the accuracy and reliability of the proposed approach.

Qualitative analysis. A visual assessment of the process was carried out by observation of the registrations across different sequences, involving 113 male and female subjects of varying body sizes and 28 different movements, with each frame of the sequence undergoing an exhaustive visual scan. Figure 4 shows several examples illustrating the results obtained from the registration process. This detailed analysis aimed to identify potential visual imperfections or important alignment errors, as well as to assess the temporal coherence in each data sequence. This approach provided a thorough understanding of the quality and consistency of the obtained results.

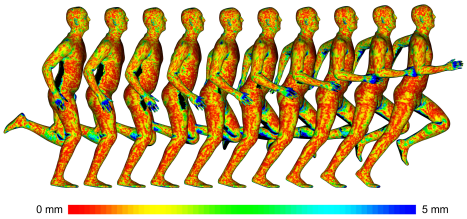


Figure 3: *Distances between the processed mesh and the raw point cloud. Occluded regions lacking raw data are represented in black.*

Figure 4: *Examples of registration of diverse subjects performing different movements.*

Temporal consistency. During the visual analysis of the 3D meshes in motion, we observed a significant reduction in jitter between frames, indicating a notable improvement in the stability and precision of three-dimensional representations over time. This decrease in jitter not only benefits the visual quality of the anima-

tions, but also increases the credibility and authenticity of the captured motion. Furthermore, it was observed that the inclusion of information from neighboring frames is essential to effectively correct pose ambiguities. By taking advantage of the contextual information provided by adjacent frames, a smoother and more natural transition is achieved between different poses, resulting in a more coherent and realistic representation of movement in temporal sequences.

6.2. Evaluation of standard containers to exchange and use virtual characters

To conduct the evaluation of the two file formats that we implemented, we created some examples of male and a female subjects in each variant of FBX and glTF formats described in section 5, namely FBX with single- and multiple-blendshapes, and glTF in Binary and ASCII. Firstly, we evaluated the capability of the format to embed the content resulting from the registration, i.e. one texture per frame and one registered mesh per frame expressed according to the model $M(s,\beta,\varepsilon_s,\theta,\varepsilon_f)$. In both cases we found similar limitations to include all the content in its original form, in particular the textures and the data models.

Textures. We were unable to animate texture. It is possible to store multiple texture images in the file, but we did not found a mechanism to transition between them. Yet, 3D modelling software such as Blender is able to cope with texture animations using a pointer to the image sequence stored in separate files. So, as an example to turn around this limitation and be able to use the full stream of images, we stored the images separately and created a Blender script to import the non-textured animated mesh with UV maps in FBX and point to the corresponding image file for each frame.

Human Body Models. We were unable to pack the soft tissue model $B(\beta,\theta)$, and to express the registered mesh animation as the pose parameters θ and ε_f. Since the formats enabled to store skinned animation using LBS plus a corrective shape per frame, we had to compute a new term ε_r per frame. This term groups the shape variation due to soft tissue deformation given by $B(\beta,\theta)$ and the ε_f term. Despite the FBX and glTF containers cannot embed these models, 3D modelling software such as Blender is able to cope with linear shape and pose-dependent corrections. Other authors [Mes23] have addressed the storage and use of linear body data. They propose to store each shape component and each pose-dependent shape corrective component as blendshapes in a FBX file and then provide a specific script for each platform (i.e. Blender, Maya, Unreal and Unity) to be load and use it. However, this approach remains insufficient since it requires additional development and scripting to fully exploit the content issued by the proposed technology.

Secondly, we tested the use of the different FBX and glTF variants into commonly used 3D viewers, 3D modelling software, 3D garment design, game engines and 3D platforms by checking if they were able to import and reproduce the registered mesh animations (i.e. skinned animation plus corrective blendshapes) with texture. We did not tested unofficial plugins or scripts available that are aimed to enable or improve the import of these files.

Importing assets in commercial 3D software. We found a wide variety of results depending on the software used (Table 1). Generally, FBX import is supported by most of the software while glTF support is less common. We also noted that claiming that a software supports FBX or glTF import does not imply that it is able to reproduce skinned animation plus blendshapes. We found no differences when importing glTF in either of its variants. Finally, between the two FBX variants, multiple blendshapes seemed to be the most commonly supported one.

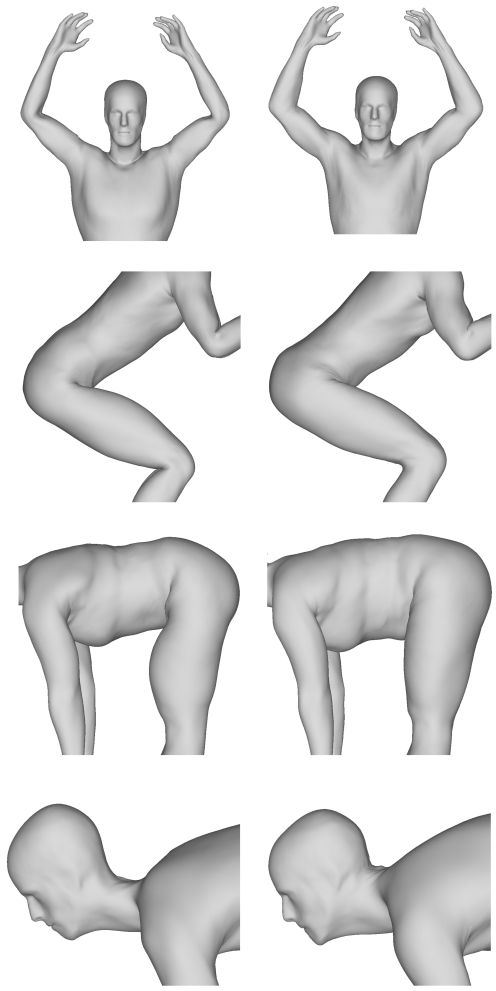
Software	FBX multi	FBX single	glTF
FBX Review	partial	partial	-
3D Viewer	partial	partial	yes
Blender	yes	no	yes
Maya	yes	partial	-
3ds Max	yes	yes	-
Unity	yes	yes	yes
Unreal Engine	yes	partial	no
Omniverse	partial	partial	partial
CLO3D	yes	partial	-

Table 1: *Summary of compatibility results of the asset import test: 'yes' indicates the ability to reproduce skinned and blendshape animations; 'no' indicates inability to reproduce neither skinned animation nor blendshape animation; 'partial' indicates ability to reproduce skinned animation but not blendshape animation; '-' indicates not supported format.*

7. Conclusions

This paper proposes an end-to-end process from the capture of the human subjects performing a real action to the delivery of animated virtual characters including texture, skinned animation and corrective blendshapes in FBX or glTF formats. It is more efficient than the current practices described in section 2 because our integrated pipeline is fully automatic, it does not depend on user skills and it does not require the use of third party software.

Our approach enables to capture 3D body surfaces of real people in action with high detail including facial gestures and hands motion and with texture. The resulting 3D data in motion is richer (i.e. over 4 million points and a set of high resolution images) than current practices such as animating static 3D scans using motion capture data or directly estimating shape and motion just from motion data [LMB14]. Moreover, compared to other accurate motion capture systems using sparse optical markers [Vic], our approach can help to save time with actors since it does not require instrumentation with physical markers. Compared to other markerless approaches [The; DAR], ours is much more comprehensive because it provides facial gestures and a more complete skeleton including spine and hands. Since it is an actual capture, our approach can confer to the virtual characters a more accurate surface representation and more natural motion and behavior than model-based approaches that just capture sparse motion data. Figure 5 illustrates the differences at different body parts (i.e. chest, arm, elbow, belly, buttock, groin, thigh, knee and neck) between animating or posing a registered mesh of a static A-Pose using LBS and the actual registration using our methods.



(a) (b)

Figure 5: *Visual comparison of identical poses of (a) 3D scans registered in A-Pose and reposed using LBS and (b) the registered meshes following our approach.*

The proposed hardware approach can achieve high raw data transfer rates making it possible to verify the point cloud captures almost in real-time and make fast decisions about the need for acquiring additional movements during the capturing sessions. Moreover, the number of sensors and their position, orientation and optics as well as camera parameters can be configured to reduce occlusions and achieve a higher level of accuracy and texture resolution for virtual simulation applications such as fluid dynamics, virtual fit or other mechanical simulations.

The automatic registration process is very robust and accurate. Thanks to the use of advanced models for pose-dependent shape corrections, the temporal consistency algorithms and the combined use of 3D points and images, it can produce realistic shape results with smooth and artifact-free motion even in frames with highly occluded body parts.

The resulting homologous mesh and the common parameterization of the skeleton and UV map provide a series of practical advantages for the enrichment of the characters in specialized software. It is possible to swap motions, postures or texture just by swapping the skeletons or texture files between different characters. Having a common rest pose (i.e. zero value for all angles) for all characters and captures facilitates pose manipulation and retargeting. Secondly, the change of topology, UV map or skeletal structure to another one that is required or preferred for the specific application is facilitated because it can be easily achieved in all the characters and motions by establishing a mapping between the two objects that is just computed once. It can be computed based on a single example or in a subset of data. These properties also simplify significantly the conduction of intra- and inter-subject analyses based on body surface or skeletal data in virtual simulation applications.

The possibility of saving the registered mesh sequences in standard file exchange formats such as FBX and glTF is also very practical to further develop the characters in 3D modelling software or to directly integrate them into 3D development platforms.

We propose an end-to-end pipeline to capture and register 3D humans in motion which is very efficient using affordable computing resources. Within our approach, since the capture of the sequences is very fast and the most time- and resource-consuming part is the processing of the scans, large-scale productions or projects that may require it could parallelize this latter operation by using more than one computer to process the captures as they are available.

This efficiency makes it particularly suitable for gathering large datasets of real humans in action which has a great potential to foster the learning of human models. Actually, we gathered different datasets and used them to learn several of the pre-calculated body models described in section 4.1. This pipeline can also be used to learn on-the-fly soft-tissue models $B(\beta,\theta)$ optimized for a particular subject or actor performing a variety of movements. Such individual DHM would move, behave and interact in a much more realistic way than using pre-calculated models for new animations that are similar to the movements used to learn the model. In a similar fashion, they can also be used to learn other pre-calculated soft-tissue models $B(\beta,\theta)$ for particular gestures, activities, sports or combat disciplines in combination or not with particular body types related those movements. Another dataset including 107725 poses from 40 adult subjects performing over 20 sports motions was produced with our technology and it is being used other authors to learn Digital Human Models for the virtual design of sports gear and garments [PLB*23].

8. Limitations and further Works

One limitation of the proposed homologous processing is that it requires that there are no other objects (or subjects) in the scene that can be confused as a part of the body and thus registered as such. To overcome this limitation, segmentation algorithms and DL models (either point-based, image-based or a combination) could be used to segment the point cloud into different subjects and objects.

Our processing also requires that the subjects/actors are minimally dressed. For clothed subjects/actors, we are exploring to extend our approach to estimate the body in motion under the clothes

using prior knowledge of the subject (e.g. shape, pose and soft-tissue) learnt from registered captures of the same subject captured with minimal clothing.

The proposed solution does not take into account the physical interaction with objects, which is particularly relevant for virtual product design and simulation. The registered meshes can be used as input for different physical simulations but it also could be interesting to explore the integration of recent research in the field of physical simulation modelling [RRTO23b; RRTO23a].

At the the current state of development of FBX and glTF standard formats, final users have to reuse, adapt or develop specific scripts for each software or platform to fully exploit the content issued by the proposed technology. In this regard, we are assessing the possibility of creating specific scripts to the most common software and platforms to take advantage of the full texture stream captured as well as to include soft tissue models. We also plan to extend our algorithms to provide USD files.

9. Acknowledgments

This research was partially funded by the Instituto Valenciano de Competitividad Empresarial (IVACE) and Valencian Regional Government (GVA) (IMAMCA/2024) and the Consellería de Innovación, Industria, Comercio y Turismo through the project CONV24/DGINN, Grants for Technology Centers for Innovation Projects in collaboration with companies, within the Smart Specialization a approach (S3). This publication is funded by the European Union (project INNDIH; GA101083002), with Ministerio de Industria, Turismo y Comercio and Valencian Regional Government (GVA).

References

[ACP03] ALLEN, BRETT, CURLESS, BRIAN, and POPOVIĆ, ZORAN. "The space of human body shapes: reconstruction and parameterization from range scans". *ACM Trans. Graph.* 22.3 (July 2003), 587–594. ISSN: 0730-0301. DOI: 10.1145/882262.882311. URL: https://doi.org/10.1145/882262.882311 2, 3.

[Ado] ADOBE. *Substance 3D Painter*. Version 9.1. URL: https://www.adobe.com/products/substance3d-painter.html 2.

[AJGA23] AL-ANSI, ABDULLAH M., JABOOB, MOHAMMED, GARAD, ASKAR, and AL-ANSI, AHMED. "Analyzing augmented reality (AR) and virtual reality (VR) recent development in education". *Social Sciences & Humanities Open* 8.1 (Jan. 2023), 100532. ISSN: 2590-2911. DOI: 10.1016/j.ssaho.2023.100532. URL: https://www.sciencedirect.com/science/article/pii/S2590291123001377 (visited on 02/05/2024) 1.

[ASK*05] ANGUELOV, DRAGOMIR, SRINIVASAN, PRAVEEN, KOLLER, DAPHNE, et al. "SCAPE: shape completion and animation of people". *ACM Trans. Graph.* 24.3 (July 2005), 408–416. ISSN: 0730-0301. DOI: 10.1145/1073204.1073207. URL: https://doi.org/10.1145/1073204.1073207 3.

[AST15] ASTM INTERNATIONAL. *ASTM D5219-15 Standard Terminology Relating to Body Dimensions for Apparel Sizing*. English. ASTM Standard. 2015. URL: https://www.astm.org/d5219-15.html 7.

[Auta] AUTODESK, INC. *Autodesk 3ds Max*. Version 2015. URL: https://www.autodesk.es/products/3ds-max/overview 2.

[Autb] AUTODESK, INC. *Filmbox*. Version 2020.1. URL: https://www.autodesk.com/products/fbx 2.

[Autc] AUTODESK, INC. *Maya*. Version 2020.4. URL: https://autodesk.com/maya 2.

[BGR*20] BAZAREVSKY, VALENTIN, GRISHCHENKO, IVAN, RAVEENDRAN, KARTHIK, et al. "BlazePose: On-device Real-time Body Pose tracking". *CoRR* abs/2006.10204 (2020). arXiv: 2006.10204. URL: https://arxiv.org/abs/2006.10204 4.

[Ble] BLENDER ONLINE COMMUNITY. *Blender - a 3D modelling and rendering package*. Version 3.6. Stichting Blender Foundation, Amsterdam: Blender Foundation. URL: http://www.blender.org 2.

[BMR23] BACHILLER, CARMEN, MONZO, JOSE M., and REY, BEATRIZ. "Augmented and Virtual Reality to Enhance the Didactical Experience of Technological Heritage Museums". en. *Applied Sciences* 13.6 (Jan. 2023). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, 3539. ISSN: 2076-3417. DOI: 10.3390/app13063539. URL: https://www.mdpi.com/2076-3417/13/6/3539 (visited on 02/05/2024) 1.

[BP07] BARAN, ILYA and POPOVIĆ, JOVAN. "Automatic rigging and animation of 3D characters". *ACM Transactions on Graphics* 26.3 (July 2007), 72–es. ISSN: 0730-0301. DOI: 10.1145/1276377.1276467. URL: https://dl.acm.org/doi/10.1145/1276377.1276467 (visited on 02/13/2024) 2.

[BPU*14] BALLESTER, ALFREDO, PARRILLA, EDUARDO, URIEL, JORDI, et al. "3D-based resources fostering the analysis, use, and exploitation of available body anthropometric data". *5th international conference on 3D body scanning technologies*. 2014, 237–247 7.

[BVN*15] BALLESTER, ALFREDO, VALERO, MARTA, NÁCHER, BEATRIZ, et al. "3D Body Databases of the Spanish Population and its Application to the Apparel Industry". *Proc. of 6th Int. Conf. on 3D Body Scanning Technologies*. 2015. URL: https://api.semanticscholar.org/CorpusID:415927 3.

[CLO] CLO VIRTUAL FASHION, INC. *Marvelous Designer*. Version 12.2. URL: https://marvelousdesigner.com 2.

[CPWD21] CHALMERS, ALAN, PARKINS, JOSEPH, WEBB, MARK, and DEBATTISTA, KURT. "Realistic Humans in Virtual Cultural Heritage". en. *Emerging Technologies and the Digital Transformation of Museums and Heritage Sites*. Ed. by SHEHADE, MARIA and STYLIANOU-LAMBERT, THEOPISTI. Communications in Computer and Information Science. Cham: Springer International Publishing, 2021, 156–165. ISBN: 978-3-030-83647-4. DOI: 10.1007/978-3-030-83647-4_11 1.

[DAR] DARI MOTION, INC. *Captury*. Version 2023.1. URL: https://captury.com 8.

[DMT12] DALIBARD, SÉBASTIEN, MAGNENAT-TALMANN, NADIA, and THALMANN, DANIEL. "Anthropomorphism of Artificial Agents: A Comparative Survey of Expressive Design and Motion of Virtual Characters and Social Robots". *Workshop on Autonomous Social Robots and Virtual Humans at the 25th Annual Conference on Computer Animation and Social Agents (CASA 2012)*. Singapore, Singapore, May 2012. URL: https://hal.science/hal-00732763 (visited on 02/13/2024) 1.

[DT13] DAANEN, H. A. M. and TER HAAR, F. B. "3D whole body scanners revisited". *Displays* 34.4 (Oct. 2013), 270–275. ISSN: 0141-9382. DOI: 10.1016/j.displa.2013.08.011. URL: https://www.sciencedirect.com/science/article/pii/S014193821300070X (visited on 02/13/2024) 2.

[DTSS07] DE AGUIAR, EDILSON, THEOBALT, CHRISTIAN, STOLL, CARSTEN, and SEIDEL, HANS-PETER. "Rapid Animation of Laser-scanned Humans". en. *2007 IEEE Virtual Reality Conference*. Charlotte, NC, USA: IEEE, Mar. 2007, 223–226. ISBN: 978-1-4244-0905-1. DOI: 10.1109/VR.2007.352486. URL: https://ieeexplore.ieee.org/document/4161028/ (visited on 02/13/2024) 2.

[DvdW98] DAANEN, HEINA. M. and van de WATER, G. JEROEN. "Whole body scanners". *Displays* 19.3 (Nov. 1998), 111–120. ISSN: 0141-9382. DOI: 10.1016/S0141-9382(98)00034-1. URL: https://www.sciencedirect.com/science/article/pii/S0141938298000341 (visited on 02/13/2024) 2.

[DVP*22] De Rosario, H., Vivas-Broseta, M.J., Pitarch-Corresa, S., et al. "Improvement of joint reaction forces and moments calculation during a step up and over task using a 4D scanner data". *Gait & Posture* 97 (2022). ESMAC 2022 Abstracts, S262–S263. ISSN: 0966-6362. DOI: https://doi.org/10.1016/j.gaitpost.2022.07.159. URL: https://www.sciencedirect.com/science/article/pii/S0966636222003587 7.

[Epi] Epic Games. *Unreal Engine*. Version 5.3.2. URL: https://www.unrealengine.com 2.

[Fac] Faceform LLC. *Faceform*. Version 2023.06. URL: https://faceform.com/ 2.

[GAK*20] Grishchenko, Ivan, Ablavatski, Artsiom, Kartynnik, Yury, et al. "Attention Mesh: High-fidelity Face Mesh Prediction in Real-time". *CoRR* abs/2006.10962 (2020). arXiv: 2006.10962. URL: https://arxiv.org/abs/2006.10962 4.

[Ghe98] Ghee, Steve. "The Virtues of Virtual Products". *Mechanical Engineering* 120.06 (June 1998), 60–63. ISSN: 0025-6501. DOI: 10.1115/1.1998-JUN-1. URL: https://doi.org/10.1115/1.1998-JUN-1 (visited on 02/13/2024) 1.

[GMB*18] Gerig, Thomas, Morel-Forster, Andreas, Blumer, Clemens, et al. "Morphable Face Models - An Open Framework". *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 2018, 75–82. DOI: 10.1109/FG.2018.00021 4.

[HLRB12] Hirshberg, David A., Loper, Matthew, Rachlin, Eric, and Black, Michael J. "Coregistration: Simultaneous Alignment and Modeling of Articulated 3D Shape". *Computer Vision – ECCV 2012*. Ed. by Fitzgibbon, Andrew, Lazebnik, Svetlana, Perona, Pietro, et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 242–255. ISBN: 978-3-642-33783-3 2, 3.

[Int17a] International Organization for Standardization. *ISO 7250-1:2017 Basic human body measurements for technological design – Part 1: Body measurement definitions and landmarks*. English. ISO Standard. 2017. URL: https://www.iso.org/standard/65246.html 7.

[Int17b] International Organization for Standardization. *ISO 8559-1:2017 Size designation of clothes – Part 1: Anthropometric definitions for body measurement*. English. ISO Standard. 2017. URL: https://www.iso.org/standard/61686.html 7.

[Khra] Khronos Group. *glTF Validator*. Version 2.0.0. URL: https://github.khronos.org/glTF-Validator 2.

[Khrb] Khronos Group. *Graphics Library Transmission Format*. Version 2.0. URL: https://khronos.org/glTF 2.

[LJLS19] Lee, Bokyung, Jin, Taeil, Lee, Sung-Hee, and Saakes, Daniel. "SmartManikin: Virtual Humans with Agency for Design Tools". *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, May 2019, 1–13. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300814. URL: https://dl.acm.org/doi/10.1145/3290605.3300814 (visited on 02/13/2024) 1.

[LMB14] Loper, Matthew M., Mahmood, Naureen, and Black, Michael J. "MoSh: Motion and Shape Capture from Sparse Markers". *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 33.6 (Nov. 2014), 220:1–220:13. DOI: 10.1145/2661229.2661273. URL: http://doi.acm.org/10.1145/2661229.2661273 8.

[LMR*15] Loper, Matthew, Mahmood, Naureen, Romero, Javier, et al. "SMPL: A Skinned Multi-Person Linear Model". *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16 2, 3.

[Mes23] Meshcapade. *SMPL Blender Addon*. 2023. URL: https://github.com/Meshcapade/SMPL_blender_addon 8.

[MK00] Mochimaru, Masaaki and Kouchi, Makiko. "Statistics for 3D Human Body Forms". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44.38 (July 2000). Publisher: SAGE Publications Inc, 852–855. ISSN: 1071-1813. DOI: 10.1177/154193120004403846. URL: https://doi.org/10.1177/154193120004403846 (visited on 03/22/2024) 2.

[Mor98] Morrissey, Mark. "Human-Centric Design". *Mechanical Engineering* 120.07 (July 1998), 60–62. ISSN: 0025-6501. DOI: 10.1115/1.1998-JUL-2. URL: https://doi.org/10.1115/1.1998-JUL-2 (visited on 02/13/2024) 1.

[MT05] Magnenat-Thalmann, Nadia and Thalmann, Daniel. "Virtual humans: thirty years of research, what next?" en. *The Visual Computer* 21.12 (Dec. 2005), 997–1015. ISSN: 1432-2315. DOI: 10.1007/s00371-005-0363-6. URL: https://doi.org/10.1007/s00371-005-0363-6 (visited on 02/13/2024) 1.

[NVI] NVIDIA Corporation. *Omniverse*. Version 2022.3.3. URL: https://www.nvidia.com/en-us/omniverse 2.

[PBP*19] Parrilla, Eduardo, Ballester, Alfredo, Parra, Francisco, et al. "MOVE 4D: Accurate High-Speed 3D Body Models in Motion". *Proceedings of 3DBODY.TECH 2019 - 10th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, Lugano, Switzerland, 22-23 Oct. 2019* (2019). URL: https://api.semanticscholar.org/CorpusID:209090065 2.

[PDL*23] Pitarch-Corresa, Salvador, De Rosario - Martínez, Helios, López - Pascual, Juan, et al. "Innovative use of 4D scanner for gait analysis of neurological disorders: A case study". *Gait & Posture* 106 (2023). ESMAC 2023 Abstracts, S166–S167. ISSN: 0966-6362. DOI: https://doi.org/10.1016/j.gaitpost.2023.07.200. URL: https://www.sciencedirect.com/science/article/pii/S0966636223012201 7.

[Pil] Pilgway. *3DCoat*. Version 2023.10. URL: https://3dcoat.com 2.

[Pixa] Pixar Animation Studios. *Universal Scene Description*. Version 23.11. URL: https://www.openusd.org/ 2.

[Pixb] Pixelmachine, SRL. *TopoGun*. Version 3. URL: https://www.topogun.com 2.

[PLB*23] Polanco, Alejandra, Lafon, Yoann, Beurier, Georges, et al. "Skinning Weights Optimization through Data-Driven Approaches: Determining the Training Dataset". Oct. 2023. DOI: 10.15221/23.29. URL: https://3dbody.tech/cap/abstracts/2023/2329polanco.html (visited on 02/09/2024) 9.

[QYSG17] Qi, Charles R, Yi, Li, Su, Hao, and Guibas, Leonidas J. "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space". *arXiv preprint arXiv:1706.02413* (2017) 4.

[RDB*22] Ruescas Nicolau, Ana V., De Rosario, Helios, Basso Della-Vedova, Fermín, et al. "Accuracy of a 3D temporal scanning system for gait analysis: Comparative with a marker-based photogrammetry system". *Gait & Posture* 97 (2022), 28–34. ISSN: 0966-6362. DOI: https://doi.org/10.1016/j.gaitpost.2022.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0966636222001904 7.

[RDBJ24] Ruescas-Nicolau, Ana V., De Rosario, Helios, Bernabé, Eduardo Parrilla, and Juan, M.-Carmen. "Positioning errors of anatomical landmarks identified by fixed vertices in homologous meshes". *Gait & Posture* 108 (2024), 215–221. ISSN: 0966-6362. DOI: https://doi.org/10.1016/j.gaitpost.2023.11.024. URL: https://www.sciencedirect.com/science/article/pii/S0966636223015047 7.

[RDP99] Robinette, K.M., Daanen, H., and Paquet, E. "The CAESAR project: a 3-D surface anthropometry survey". *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*. 1999, 380–386. DOI: 10.1109/IM.1999.805368 3.

[RRTO23a] RAMÓN, PABLO, ROMERO, CRISTIAN, TAPIA, JAVIER, and OTADUY, MIGUEL A. "FLSH - Friendly Library for the Simulation of Humans". Oct. 2023. DOI: 10.15221/23.20. URL: https://3dbody.tech/cap/abstracts/2023/2320ramon.html (visited on 02/09/2024) 10.

[RRTO23b] RAMÓN, PABLO, ROMERO, CRISTIAN, TAPIA, JAVIER, and OTADUY, MIGUEL A. "SFLSH: Shape-Dependent Soft-Flesh Avatars". *SIGGRAPH Asia 2023 Conference Papers*. SA '23. , Sydney, NSW, Australia, Association for Computing Machinery, 2023. DOI: 10.1145/3610548.3618242. URL: https://doi.org/10.1145/3610548.3618242 10.

[RTB17] ROMERO, JAVIER, TZIONAS, DIMITRIOS, and BLACK, MICHAEL J. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 36.6 (Nov. 2017) 4.

[The] THEIA MARKERLESS, INC. *Theia3D*. Version 2023.1. URL: https://www.theiamarkerless.ca 8.

[Uni] UNITY TECHNOLOGIES. *Unity*. Version 2022.3.19f1. URL: https://unity.com 2.

[URI*22] URIEL, JORDI, RUESCAS, A., IRANZO, SOFIA, et al. "A methodology to obtain anthropometric measurements from 4D scans". *Proceedings of the 7th International Digital Human Modeling Symposium 7(1): 12*. Aug. 2022, 13. DOI: 10.17077/dhm.31758 7.

[Vic] VICON MOTION SYSTEMS, LTD. *Vicon*. URL: https://www.vicon.com 8.

[WC95] WU, GE and CAVANAGH, PETER R. "ISB recommendations for standardization in the reporting of kinematic data". *Journal of Biomechanics* 28.10 (1995), 1257–1261. ISSN: 0021-9290. DOI: https://doi.org/10.1016/0021-9290(95)00017-C. URL: https://www.sciencedirect.com/science/article/pii/002192909500017C 3.

[WFD*22] WOLF, ERIK, FIEDLER, MARIE LUISA, DÖLLINGER, NINA, et al. "Exploring Presence, Avatar Embodiment, and Body Perception with a Holographic Augmented Reality Mirror". *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. ISSN: 2642-5254. Mar. 2022, 350–359. DOI: 10.1109/VR51125.2022.00054. URL: https://ieeexplore.ieee.org/document/9756775 (visited on 02/07/2024) 1.

[WMG14] WAECHTER, MICHAEL, MOEHRLE, NILS, and GOESELE, MICHAEL. "Let There Be Color! — Large-Scale Texturing of 3D Reconstructions". *Proceedings of the European Conference on Computer Vision*. Springer, 2014 6.

[WSA*02] WU, GE, SIEGLER, SORIN, ALLARD, PAUL, et al. "ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine". *Journal of Biomechanics* 35.4 (2002), 543–548. ISSN: 0021-9290. DOI: https://doi.org/10.1016/S0021-9290(01)00222-6. URL: https://www.sciencedirect.com/science/article/pii/S0021929001002226 3.

[WvD*05] WU, GE, VAN DER HELM, FRANS C.T., (DIRKJAN) VEEGER, H.E.J., et al. "ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand". *Journal of Biomechanics* 38.5 (2005), 981–992. ISSN: 0021-9290. DOI: https://doi.org/10.1016/j.jbiomech.2004.05.042. URL: https://www.sciencedirect.com/science/article/pii/S002192900400301X 3.

[XBZ*20] XU, HONGYI, BAZAVAN, EDUARD GABRIEL, ZANFIR, ANDREI, et al. "GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Oral)*. 2020, 6184–6193. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Xu_GHUM__GHUML_Generative_3D_Human_Shape_and_Articulated_Pose_CVPR_2020_paper.html 2, 3.

[YZF*24] YANG, YANHONG, ZHANG, HAOZHENG, FERNÁNDEZ, ALFREDO BALLESTER, et al. "Digitalization of Three-Dimensional Human Bodies: A Survey". *IEEE Transactions on Consumer Electronics* (2024), 1–1. DOI: 10.1109/TCE.2024.3363616 7.

[ZBV*20] ZHANG, FAN, BAZAREVSKY, VALENTIN, VAKUNOV, ANDREY, et al. "MediaPipe Hands: On-device Real-time Hand Tracking". *CoRR* abs/2006.10214 (2020). arXiv: 2006.10214. URL: https://arxiv.org/abs/2006.10214 4.