Three-step Alignment Approach for Fitting a Normalized Mask of a Person Rotating in A-Pose or T-Pose Essential for 3D Reconstruction based on 2D Images and CGI Derived Reference Target Pose

Gerald Adam Zwettler[1,2][a], Christoph Praschl[1][b], David Baumgartner[1][c], Tobias Zucali[3][d],
Dora Turk[3][e], Martin Hanreich[1] and Andreas Schuler[1,4][f]

[1]*Research Group Advanced Information Systems and Technology (AIST), University of Applied Sciences Upper Austria,*
Softwarepark 11, 4232 Hagenberg, Austria
[2]*Department of Software Engineering, School of Informatics, Communications and Media,*
University of Applied Sciences Upper Austria, Softwarepark 11, 4232 Hagenberg, Austria
[3]*AMB GmbH, amb-technology.ai, Hafenstraße 47-51 4020 Linz, Austria*
[4]*Department of Bio and Medical Informatics, School of Informatics, Communications and Media,*
University of Applied Sciences Upper Austria, Softwarepark 11, 4232 Hagenberg, Austria

Keywords: Elastic Shape Alignment, Human Body Pose Detection, 3D Body Reconstruction, Silhouette Reconstruction.

Abstract: The 3D silhouette reconstruction of a human body rotating in front of a monocular camera system is a very challenging task due to elastic deformation and positional mismatch from body motion. Nevertheless, knowledge of the 3D body shape is a key information for precise determination of one's clothing sizes, e.g. for precise shopping to reduce the number of return shipments in online retail. In this paper a novel three step alignment process is presented, utilizing As-Rigid-As-Possible (ARAP) transformations to normalize the body joint skeleton derived from OpenPose with a CGI rendered reference model in A- or T-pose. With further distance-map accelerated registration steps, positional mismatches and inaccuracies from the OpenPose joint estimation are compensated thus allowing for 3D silhouette reconstruction of a moving and elastic object without the need for sophisticated statistical shape models. Tests on both, artificial and real-world data, generally proof the practicability of this approach with all three alignment/registration steps essential and adequate for 3D silhouette reconstruction data normalization.

1 INTRODUCTION

To accurately determine one's clothing sizes, conventional manual measurement at the tailor or shop are commonly used, as well as 3D scans of the entire body. To allow for manual measurements at home, e.g. to prevent from bad purchase of not fitting clothes in online retail, depth sensor based measurement are hardly applicable. Many customers only have a rough knowledge of their individual body size. Additionally, clothing sizes are neither nationally nor internationally standardized and different target markets and manufacturers have developed different size names and standards. The annual costs for return shipments in the online retail market are a significant cost factor, numbered with approximately 550 billion USD for the U.S. market in year 2020 (Orendorfff, 2019) (Mazareanu, 2020). While on average around 12-15% returns are common in online trade, returns in the clothing sector reach a level of just under a third (Statista, 2012) and mismatch in the size is the biggest culprit (Murphy, 2020). Since returned goods always cause effort and loss of value, there are organizational efforts to reduce this high proportion.

In the course of the research project TrueSize S.M.B.S., the technical foundation for an innovative smart mobile body scanner utilizing monocular video feed is therefore explored. With accurate knowledge of the own clothing size, the costs for online clothing retailers can be reduced.

[a] https://orcid.org/0000-0002-4966-6853
[b] https://orcid.org/0000-0002-9711-4818
[c] https://orcid.org/0000-0002-0189-4718
[d] https://orcid.org/0000-0003-3771-0041
[e] https://orcid.org/0000-0003-4348-5426
[f] https://orcid.org/0000-0003-1074-3222

Zwettler, G., Praschl, C., Baumgartner, D., Zucali, T., Turk, D., Hanreich, M. and Schuler, A.
Three-step Alignment Approach for Fitting a Normalized Mask of a Person Rotating in A-Pose or T-Pose Essential for 3D Reconstruction based on 2D Images and CGI Derived Reference Target Pose.
DOI: 10.5220/0010194102810292
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 281-292
ISBN: 978-989-758-488-6

1.1 Problem Statement

Scanning a human person with a monocular smart phone camera to reconstruct precise 3D avatars is a challenging task in computer vision due to the highly dynamic nature of the environment, starting with the issue of object localization, segmentation and estimation of the orientation. To gain sufficient 3D information of an object with only one camera, incorporating several views is inevitable. Thus, for 3D reconstruction of an object with only one camera system, either the object needs to rotate with static camera or the camera is moved around the statically placed object. This constraint holds for both, monocular RGB video feed and RGB-D depth data. Additional problems with reconstruction from monocular images generally are an inaccurate camera model, non-orthogonal aligned object/camera and the unknown distance to the object preventing a direct calculation of the pixel to mm scale factor for real-world proportions and measurements. Furthermore, if a human is rotating on a spot being captured with the video sequence, additional problems like inner body movement due to the 26 most prominent joints of human body kinematic chain arise (Zatsiorsky, 1998). Besides rotating on a spot, the local rotation axis cannot be kept that precisely, which indicates that straight forward silhouette reconstruction is impossible.

1.2 State of the Art

The 3D reconstruction of objects from monocular images requires different state of the art approaches in computer vision.

Localization of humans in images can be achieved utilizing HOG features (Liu et al., 2013) or Haar Cascades (Aguilar et al., 2017) but recent improvements in deep learning (DL) allow for advancements. Nowadays, precise localization of humans can be achieved utilizing DL models, e.g. OpenPose (Cao et al., 2019) approximating limb and bone poses in 2D even in case of partial occlusions. Based on this limb approximation, estimation of the orientation becomes feasible too, assessing the hip-rotation with a plane between chest and left/right hip (Wei et al., 2019) or utilizing specific DL models (Xiang et al., 2017).

For person segmentation in videos, Grab cut approaches (Yu et al., 2019) as well as utilizing deep learning approaches (Liu and Stathaki, 2017) is applicable. With static camera systems and constant lighting, modeling the background as foundation for subtraction/differential imaging becomes possible as well as e.g. evaluated in (Zeng et al., 2018) but is generally very sensitive to the signal-to-noise-ratio, motion and changes on the lighting condition. The domain of person segmentation can also be addressed utilizing 3D statistical shape models (Cootes et al., 1995; Pishchulin et al., 2017) with demand for dataset-specific registration and model adaption based on 2D model projection. Other skeleton-based approaches for 2D and 3D shape alignment originate from the computer graphics and animation domain (Song et al., 2013).

In the area of 3D shape reconstruction from monocular images utilizing several camera systems, silhouette reconstruction is applicable (Mulayim et al., 2003) to derive the visual hull of the target object applying voxel carving on discrete 3D volume. If some kind of depth approximation is present, silhouette reconstruction can be refined compared to visual hull (Lin and Wu, 2008). Depth information thereby can result from ToF sensor, DL models, stereo image matching or utilizing photo-consistency on computer graphics lighting models for reverse engineering for depth approximation from input intensity images (Seitz and Dyer, 1999). Furthermore, for skeleton-based reconstruction approaches, epipolar geometry allows to derive or validate 3D joint positions of the person from the 2D images (Li et al., 2018).

1.3 Related Work

For 3D reconstruction of static objects, the use of 3D sensors allows the implementation of holistic scan solutions. With a moving scanner applying local surface registration for building up an entire 3D mesh model, ReconstructME offers a comprehensive framework (Heindl et al., 2015). Even a Kinect sensor could be used for 3D reconstruction process (Marchal and Lygren, 2017). With only one static camera system incorporated, rotation of the person leads to best results (Fit3D, 2020) while placing several camera systems in a box, the person just has to stay steady for a couple of seconds (Sizestream, 2020). Nevertheless, all of these approaches necessitate a supporting person and special hardware not present when trying to measure at home.

Using statistical shape models for human body anatomy together with registration and fitting processes, it becomes the first time possible to get a 3D reconstruction from a single monocular smart phone feed only (Kocabas et al., 2019). Nevertheless, these accurately fitted models are still only a generic shape not considering anatomical variability.

To detect and analyze a human being in RGB input images, both skeleton and body part segmentation are a key aspect (Varol et al., 2017; Fang et al., 2018) interpreting pose detection as alignment task. In the work of Varol et al. (Varol et al., 2017) input skeletons are used to register and align a 3D statistical human body shape model. Thereby, the CAESAR-dataset derived statistical shape models are adjusted to the input pose to implicitly allow for body part segmentation and human 3D pose estimation as an implicit 3D body reconstruction. Based on this 3D model alignment process, an arbitrary amount of synthesized CGI images in real-world pose is provided, thus allowing to train deep learning models for body part segmentation and relative depth approximation.

Another skeleton-based approach is proposed by Fang et al. (Fang et al., 2018). It utilizes detected skeleton keypoints to search for the closest body pose match in a database of ground truth samples, directly transferring the roughly fitting body part segmentation. A fully rigid 2D alignment process is performed to roughly align the source and target body skeleton. Filling of the holes resulting from rigid body part transformation and alignment to the input person silhouette in the image data is then achieved by a deep learning model. In contrast to this concept, we utilize ARAP transformation and the input image body silhouette to directly align the measured body to the reference target pose.

For this research work, some relevant aspects for reconstructing the body of a person as 3D mesh have already been published in the past as preliminary work. The human-computer-interaction paradigms for Hough-Space related member card analysis (Pointner et al., 2018) is a key for getting the pixel-to-world scale for accurate measurements. To improve the rotation estimation of human beings analyzed with OpenPose, the incorporation of optical flow and an elliptic body model lead to further improvement (Baumgartner et al., 2020).

1.4 Research Question

We aim to answer the following research question: *Is it possible to create a pose-normalized person mask based on a single monocular RGB image and pose information from a CGI reference avatar to allow for 3D silhouette reconstruction of a non-static and elastic object?* While targeting this issue, this paper makes the following contributions:

- Approach for segmenting a person in a RGB image using pose information with basic anthropomorphic constraints.
- CGI based 2D target pose as reference image in the alignment process.
- 2D Shape alignment of person masks with three-step alignment/registration approach including As-Rigid-As-Possible (ARAP) transformations by aligned target skeleton-points.

As proof of concept, the applicability of silhouette reconstruction on the aligned person views is evaluated.

1.5 Alignment Approach Overview

In our proposed alignment approach we utilize OpenPose (Cao et al., 2019) to determine the body skeleton from a person rotating in front of a monocular camera system. The body silhouette is thereby segmented with Grab cut (Rother et al., 2004) using seed areas derived from a generic anthropomorphic model. After estimating the body orientation by analyzing the optical flow (Baumgartner et al., 2020), a reference CGI avatar is utilized to align the skeletal model via the ARAP transformation (Sorkine and Alexa, 2007) following the mesh triangulation. With subsequently to follow distance-map based rigid registration w.r.t. the CGI reference contours and a refinement process for the body extremities (arms, legs), the normalized body mask allows 3D silhouette reconstruction of an elastic object, namely the person rotating.

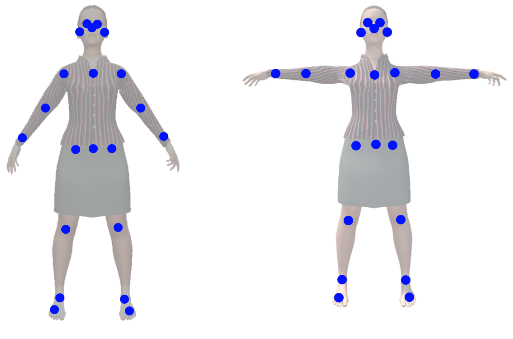
2 MATERIAL

In the following sections we present our method for creating pose-normalized person masks from a monocular RGB image. The shown examples are based on CGI models of the open source software MakeHuman (Bastioni et al., 2008), as shown in Figure 1. In addition to the computer generated inputs, we also use real world images to evaluate our alignment pipeline. Therefore, we have acquired images in portrait orientation utilizing a static camera system that are resampled to 800×1000 pixel. To fully capture the person's body, eight images at equidistant rotation of 45° each are acquired.

Furthermore, we use OpenPose (Cao et al., 2019) to retrieve a Body25 key point model (Hidalgo and Fujisaka, 2019) representation of the person's pose. Our model is simplified in small details in the foot area, where we only utilize the pose information of toes as group and using the ankle without heels. Similarily, the skeleton of the head is simplified by connecting head's main point with the eyes, see Figure 1 for our proposed adaption of the pose model.

3 METHODOLOGY

In this chapter we present our approach for calculating a pose-normalized person mask from a monoc-



(a) (b)

Figure 1: Reference pose-model in A-pose (a) and respectively in T-pose (b) used as basis for the transformation process. The female 3D model in the background is just displayed for a better understanding of the pose model.

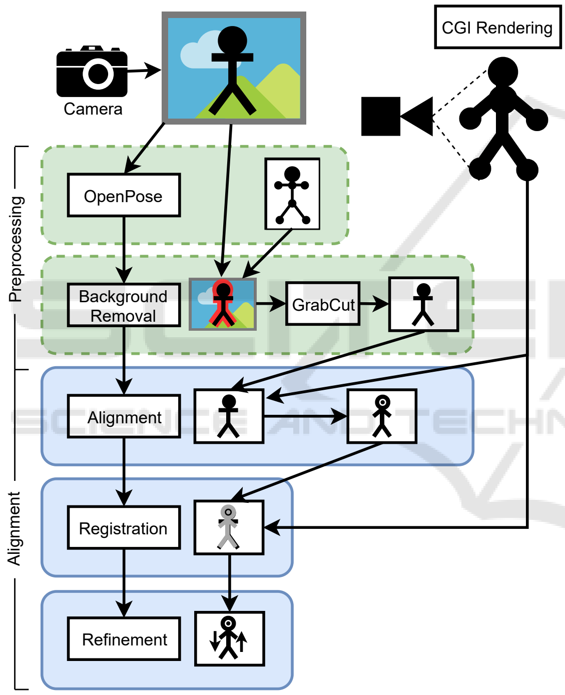


Figure 2: Overview of the process with a monocular image as input, from which pose information is extracted, that is in turn used for the background removal. Based on the created person mask the alignment is done using a reference pose from a CGI model. The alignment result is finally registered and refined.

ular RGB image and known target pose information derived from reference CGI avatars. This method is based on two key processes, the pre-processing as person segmentation (OpenPose, GrabCut) and the actual pose-normalization (Alignment, Registration, Refinement) as three-step alignment pipeline for body mask normalization, see Figure 2.

3.1 Object Segmentation

The approach we have chosen to follow is applying GrabCut (Rother et al., 2004), an optimization algorithm that estimates a Gaussian mixture model of the color distribution associated with the target and the color distribution associated with the background based on graph cuts, initialized in a novel anthropomorphic way. Namely, instead of adopting a user-specified bounding-box surrounding the target as the initial "guess" for the foreground (FG) and background (BG), we form a mask image that incorporates more information about the object and is based upon the skeletal points detected via OpenPose. The complete set of the adapted `Body25` key point model is used here to maximize the detection accuracy. The interference with the algorithm as introduced above is possible because the GrabCut algorithm can additionally take the seed information for the probable foreground (PR_FG) and probable background area (PR_BG) as arguments. By initializing the probable foreground and probable background we are giving the algorithm more hints as to where the person is expected to be, yet allowing it enough freedom to operate, that is, search for the optimal graph cut.

The procedure goes as follows. The detected skeletal points are firstly connected into a skeleton-like structure composed of straight lines mimicking human bones. The thickness of a bone is selected to be proportional to the product of the certainties the linked points are estimated with, provided by OpenPose. This skeleton-like composition is assumed to surely belong to the human that is to be segmented, i.e. foreground (FG). The layer that follows second is the probable foreground (PR_FG) and its creation is more intricate. The probable foreground should capture as much of the human body as possible, but not more than that, as it might steer the optimization in a wrong direction in case the background scenery is lively. The reasons why the process of defining it is intricate lie in the fact that the information about the shape of the human body is only partially contained in the skeleton points. An anthropomorphic model of the human anatomy would therefore be needed together with categorical references specific to the person in the image in order to complete the shape of the body. Although pursuing this concept is still our ultimate objective, our current implementation represents a compromise between the algorithm accuracy and complexity. More precisely, the probable foreground is obtained through bone-specific morphological dilation of the (sure) foreground. The same reasoning is adopted to define the probable background. Finally, the distinctively designed mask is given to the

GrabCut algorithm for refinement.

3.2 Binary Person Silhouette Image Alignment and Normalization

This section describes the second step of the presented process and refers to the pose alignment of the binary person mask. For this, a 3D avatar model is used as reference pose. This pose is adapted according to the person's original limb sizes, that are defined by the extracted skeleton points from the input image. Using those adapted skeleton joints as anchors, the person mask is transformed with an ARAP approach.

3.2.1 CGI Rendering of Reference Person

The pose normalization uses a three-dimensional CGI model as reference. This model incorporates skeleton positions and thus determines the target pose of the person in the image and can basically represent any state, for example T-pose or A-pose as shown in Figure 1. In the presented approach we are using a model in T-pose as transformation reference. Based on the reference model and the position of the required limbs, the associated skeleton joints can be determined and used in the subsequent transformation process. Since only the skeleton information is required, textures are not necessary for the model.One issue that has to be taken into account is the field of view of the camera and the projection of the reference model joints from 3D to 2D which is covered in a preprocessing step to allow for undistorted alignment.

We consider the rotation angle of the person in the image as given. Using this rotation we can project the three-dimensional model into the 2D space of the binary image that is normalized in the following steps.

3.2.2 Limb Model Alignment

The CGI reference model represents the 2D projected target pose, but is not associated with the limb sizes of the person. Thus, only the mid chest point P_{mc} of the reference model M_{ref} as well as the angle between the limbs as kinematic chains $C(P_1, P_2, ..., P_n)$ are utilized to align the input limb model M_{in} of the person. The input skeleton positions $P_i \in \mathcal{P}_{skel}$ are thereby transformed as P'_i in a way that the limb orientation is aligned according to reference model M_{ref} skeleton denoted as P''_i. Starting from the mid chest point, the kinematic chains for the head $C_{head}(P_{mc}, P_{head})$ and the shoulders $C_{shoulders_{[l|r]}}(P_{mc}, P_{shoulder_{[l|r]}})$ as well as the spine $C_{spine}(P_{mc}, P_{mh})$ and subsequently the hips as $C_{hips_{[l|r]}}(P_{mh}, P_{hip_{[l|r]}})$ are aligned. With the torso aligned, the exterior connected limbs, i.e. arms $C_{arm_{[l|r]}}(P_{shoulder_{[l|r]}}, P_{elbow_{[l|r]}}, P_{hand_{[l|r]}})$ and legs as $C_{leg_{[l|r]}}(P_{hip_{[l|r]}}, P_{knee_{[l|r]}}, P_{foot_{[l|r]}}, P_{toe_{[l|r]}})$, are aligned according to the reference pose for left and right body side cf. $[l|r]$.

Since some limbs influence other limbs, e.g. when moving the upper arm also the forearm is moved, the order is important in the alignment process. For this reason, the first step is to equate the chest and hip position of the source and the reference skeleton as $C'_{spine}(P'_{mc}, P'_{mh})$. Afterwards, all limbs are transformed starting with the limbs, that are anatomically near to the body center, namely the shoulders $C'_{shoulders_{[l|r]}} := T(C_{shoulders_{[l|r]}}, P'_{mc}, M_{ref})$, short as $C'_{spine} \rightarrow C'_{shoulders}$, and the hips as $C'_{spine} \rightarrow C'_{hips}$. Based on the transformed positions for the shoulder ($P'_{shoulder_{[l|r]}}$) and the hips ($P'_{hip_{[l|r]}}$), the left and right arms, as well as the legs, are transformed as $C'_{shoulders_{[l|r]}} \rightarrow C'_{arm_{[l|r]}}$ and $C'_{hips[l|r]} \rightarrow C'_{leg_{[l|r]}}$, respectively.

For the transformation of a kinematic chain, the limbs are iteratively transformed with the end point of the last limb as start. Thus, $T(C, P'_{start}, M_{ref}) := (P'_{start}, R(P'_{start}, P_{start}, P_2, \alpha_1), ..., R(P'_{n-1}, P_{n-1}, P_n, \alpha_{n-1}))$ for $C(P_1, P_2, ..., P_n)$.

The difference between target rotation $\alpha_i = \angle\overrightarrow{P_i, P_{i+1}}$ derived from the reference model M_{ref} and the current rotation $\beta_i = \angle\overrightarrow{P'_i, P'_{i+1}}$ is used to calculate the corrective 2D rotation matrix $R_{2D}(\phi)$ to apply, see Equation 1.

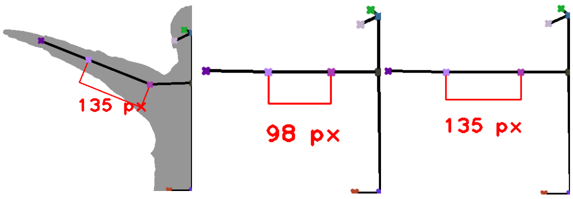
The rotation transformation R per limb is applying a rotation to balance the orientation of the particular limbs as well as a translation to fix the position w.r.t. the transformed input skeleton position P'_{start}.

$$\begin{aligned} P'_i = R(P'_{i-1}, P_{i-1}, P_i, \alpha_{i-1}) := \\ (P_i - P_{i-1}) \cdot R_{2D}(\alpha_{i-1} - \beta_{i-1}) + P'_{i-1} \end{aligned} \tag{1}$$

The limb alignment is shown using clippings in Figure 3, where (a) shows the original skeleton extracted from the input image, (b) the reference model and (c) the aligned model, with the limb orientations of (b) and the limb sizes of (a).

3.2.3 Mesh Alignment

The mesh alignment process is based on the segmented person mask, as well as the input skeleton points aligned to the reference model M'_{in}. This part of the presented method can be separated in five major steps, namely (I) contour extraction, (II) contour simplification, (III) triangulation, (IV) mesh transformation and (V) mesh projection.



(a) (b) (c)

Figure 3: Comparison of clippings of the original skeleton from the input image overlaying the segmentation result (a), the reference model (b) and the aligned model (c). The aligned model consists of the limb sizes of (a) and the limb orientation of (b).

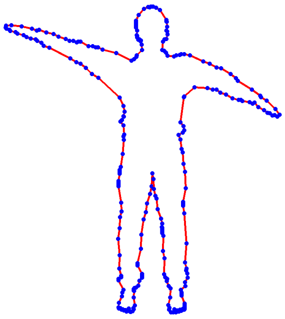


Figure 4: The largest extracted contour of the segmented image (continuous red line) is simplified to the most relevant contour points (blue) applying the Douglas-Peucker simplification algorithm with a threshold of 1.0.

First of all, the contour set $\mathcal{C} = (c_1, c_2, ..., c_n)$ with $c_i =< x_i, y_i >$ as 2D border coordinates is extracted from the segmented binary mask using the topological structural analysis algorithm of Suzuki and Abe (Suzuki et al., 1985). This algorithm extracts all continuous contours in a given binary image. We assume that the largest contour in the image represents the person after the segmentation step and all other contours represent some irrelevant segmentation fragments, see Fig. 4.

The extracted contour represents a continuous sequence of neighboring pixels. Since pixel neighbors of a contour are positioned homogeneously along a virtual polyline, those pixels can be represented in a simpler way using the extreme points along the contour, where the orientation of the pixel sequence changes with a certain threshold. For this process we represent the polygon as a connected sequence of points and use the Douglas-Peucker algorithm (Douglas and Peucker, 1973) parameterized with a tolerance of 1.0. This algorithm reduces the number of components of a curve and results in a similar shape with fewer points, see Figure 4.

In the following step, the binary mask is triangulated based on the minimized contour, in combination with the skeleton points of the input image. To keep the triangles within the body silhouette, a constrained triangulation method is required. Chew presents an adapted version of Delaunay's triangulation process (Chew, 1989) that considers non-crossing edges as borders perfectly applicable for our task. The algorithm uses the contour points as well as the skeleton points and results in a set of triangles, a so called mesh, representing the given binary mask. To improve the input for the transformation not only are we using the skeleton points but also interpolated interim positions along the limbs. In the following example images we have used $m = 5$ interim positions per limb. Using this procedure, we are able to create a denser mesh, that can be transformed in a more local way. Such a mesh is shown in Figure 5.a.

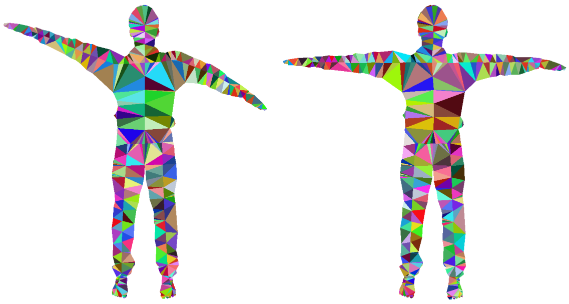


Figure 5: Triangulation result of the Constrained Delaunay algorithm using the simplified mask contour as constraints and the aligned limb model with $n = 5$ interpolated interim points per limb (left). Resulting mesh after deformation of triangulation result using the interpolated reference model's skeleton points as moving anchors for the ARAP transformation algorithm (right).

The skeleton points are added to the triangulation process for defining static anchors for the transformation. Those anchors are used as reference points for the deformation of the mesh by changing their position. The target positions are represented by the transformed input model M'_{in}. All connected mesh points are as well transformed in a weighted way based on their distance to the surrounding anchors. This process is done using the ARAP transformation algorithm (Sorkine and Alexa, 2007). The algorithm's area of application is the deformation of three-dimensional surfaces. Therefore, it uses indices of points in the original mesh together with the destination position of these points and transforms the mesh according to this movement of anchor points. In our presented approach we are using the algorithm in a 2D context, so we have to project the triangulation result into the 3D space by converting 2D pixels to 3D points with a depth-information of 0.0.

Since we consider the orientation of the person, i.e. rotation around the longitudinal axis of the body, as given, we can determine which skeleton points are visible in the image and can constraint the transformation. For example, the transformation of the mask of a person with the left half in front of the camera, should not consider points of the right half (e.g. right shoulder), because they are covered by the visible half of the body. Those points are removed from the anchor list before transforming the triangulated mesh. The result of the mesh transformation is shown in Figure 5.b.

After the transformation process we project the three-dimensional mesh back to the two-dimensional image space for re-creating the binary mask.

3.3 Registration of Multi-perspective Silhouette Images

According to the orientation of the person relative to the camera, the accuracy of OpenPose skeletal points varies, what manifests in vertically misplaced hip and shoulder positions. Thus, as not even P_{mc} can be used as reliable anchor, the aligned limbs and body silhouette must be correctly placed applying a registration process. From the CGI rendering, not only the reference skeleton/limb positions P''_i but also the reference contour $\mathcal{C}''$ is derived. As scale is already balanced with the limb alignment process, contour registration only has to cover translations T_x, T_y and rotation R as affine transformations. While the registration problem could easily be addressed utilizing ICP (iterative closest point), the use of an euclidean distance map $\mathcal{D}_{euclid}$ allows to speed up the asymptotic run-time complexity from $O(n*m)$ to $O(n)$ for $n = |\mathcal{C}|$ and $m = |\mathcal{C}''|$. The rigid registration problem is defined with mean-squared error (MSE) metric as

$$\mathcal{C}' = Trans(\mathcal{C}, \theta_{best}, T_{x_{best}}, T_{y_{best}}) \tag{2}$$

where the best transformation parameters lead to minimal squared distances between the contours $\mathcal{C}'$ and $\mathcal{C}''$, see Equation 3

$$\begin{gathered} \theta_{best}, T_{x_{best}}, T_{y_{best}} = \\ \underset{\theta, T_x, T_y}{\operatorname{argmin}} \sum_{i=1}^{|\mathcal{C}|} (\mathcal{D}_{euclid}(\mathcal{C}'')[Trans(\mathcal{C}, \theta, T_x, T_y)[i]])^2 \end{gathered} \tag{3}$$

$$\mathcal{C}'_{Trans} = Trans(\mathcal{C}, \theta_{best}, T_{x_{best}}, T_{y_{best}}) \tag{4}$$

with transformation parameters in discrete search space $\theta \in [-\theta_{min}; \theta_{max}]$, $T_x \in [-T_x; T_x]$ and $T_y \in [-T_y; T_y]$. The discrete search space thereby comprises $k = 11$ steps, i.e. radius $r = 5$, for each of the three variables (θ, T_x, T_y) and for each of the $m = 10$ optimization runs according to scale factor s_i with $[-r*s_i, -(r-1)*s_i, ..., 0, ..., (r-1)*s_i, r*s_i]$ as search offset for globally optimal parameters from the last entire run. To go over from global to local search with increasing number of optimization runs performed, the search space scale factor is reduced with $s_i = s_{i-1} * 0.9$ and initial s_1 defined from image resolution and anthropomorphic considerations. Thus, for $m = 10$ optimization runs and $r = 5$ search space radius, overall only $10*11*11*11 = 13,310$ discrete transformation parameter permutations need to be evaluated to significantly compensate for OpenPose pose inaccuracies. The transformation is applied to the skeletal points P'_i too, denoted as P'_{trans_i}.

3.4 Refined Alignment of Human Body Limbs as Kinematic Chains

Alignment of the limbs and subsequent registration as delineated in section 3.3 generally can compensate for displacements of human motion during rotating on the spot. Nevertheless, especially for thin limbs such as the arms and very slim people it is important, that the aligned contour stays precisely within the expected target limb areas as prerequisite for 3D reconstruction. Thus, the kinematic chains with skeletal points P'_{trans_i} are registered again by calculating the best correction rotation β_i at each joint position. The error is thereby determined with the target distance map $\mathcal{D}_{euclid}(\mathcal{C}'')$ on the contour points $\mathcal{C}'_{Trans}$ that are transformed according to the closest limb position $P'_{trans_j} = ClosestPoint(\mathcal{C}'_{Trans}, P'_i)$ for every skeletal point $P_i \in \mathcal{P}'$. Thus, the optimal correction rotations $\mathcal{G} = (\gamma 1, \gamma 2, ..., \gamma n)$ are searched to minimize the MSE on the contour distance map positions, see Equation 5-7.

$$\begin{gathered} \mathcal{G}_{best} = (\gamma best_1, \gamma best_2, ..., \gamma best_n) = \\ \underset{(\gamma 1, \gamma 2, ..., \gamma n)}{\operatorname{argmin}} \sum_{i=1}^{n=|\mathcal{C}|} (\mathcal{D}_{euclid}(\mathcal{C}'')[T_{corr}(\mathcal{C}'_{Trans}[i], \\ ClosestLimbP(\mathcal{C}'_{Trans}[i]), \gamma i))^2 \end{gathered} \tag{5}$$

$$\begin{gathered} T_{corr}(\mathcal{C}'_{Trans}[i], P'_i, \gamma i) := \\ (\mathcal{C}'_{Trans}[i] - P'_i) * R_{2D}(\gamma i) + P'_i \end{gathered} \tag{6}$$

$$\begin{gathered} ClosestLimbP(C_i) := P'_j | P'_j \in \mathcal{P} \wedge \\ (\forall P'_k \in \mathcal{P}, P'_j \neq P'_k : ||C_i - P'_j|| \leq ||C_i - P'_k||) \end{gathered} \tag{7}$$

The limb positions $P'_j \in \mathcal{P}$ are then transformed with the optimal gamma values $\mathcal{G}_{best} = (\gamma best_1, \gamma best_2, ..., \gamma best_n)$ as $P'''_j = T_{corr}(P'_{j-1}, P'_j, \gamma best_j)$. These transformed limb positions implicitly incorporate the limb corrections from the pose (cf. section 3.2.2), the registration (cf. section 3.3) and the recently introduced refinement. Nevertheless, the contour will not be properly transformed close to the joints as only the closest limb is taken into consideration. Consequently, the limb alignment process is re-run once with the optimized limb positions P_j''' replacing M_{ref} as new optimized target positions for transforming the original positions P_j. This way, the ARAP transformation is re-run on the optimized target positions for improved and reliable quality of results.

4 IMPLEMENTATION

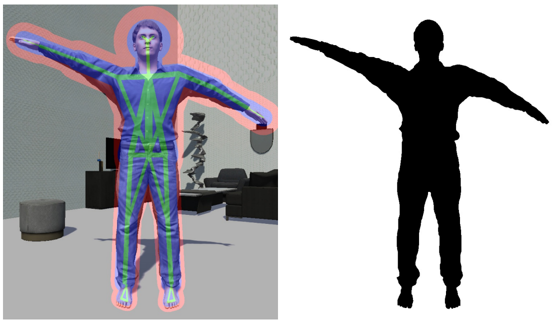
The implementation of the presented approach follows a microservice architecture and is due to the project environment mainly done using the Java programming language. While the image processing parts were developed using the Java build of the OpenCV library,(Bradski, 2000), the mesh transformation is done using the C++ libigl library,(Jacobson et al., 2018). To be able to access the C++ transformation service from the Java side, it is wrapped by a RESTful service and put into a Docker,(Merkel, 2014) container, that receives the triangulated mesh as OFF-file as well as the source and target anchor points as CSV-files and responses with an OFF-file containing the transformed mesh.

5 RESULTS

5.1 Object Segmentation

This section discloses validation of the person segmentation routine as described in section 3.1 on a CGI example with a nonuniform background.

Figure 6.a displays the considered image together with the mask designed to initialize the GrabCut algorithm laid over it. The sure foreground composed of connections between the joints detected by OpenPose is given in green. The connection are made in a way that allows us to capture the basic structure of almost any body shape. Dilation of this structure in a fashion that aims to introduce proportions common for the human body yields probable foreground given in blue. The major part of the person's body is covered adequately; nevertheless, a special attention had to be given to the area around the hands and feet as the data indicating joints there are sparse. The probable foreground is "wrapped" by the probable background indicated in red. Its thickness is a precarious parameter as everything outside of it will be discarded, and everything inside will be given a fair chance to belong to the person. This step finalizes the initial seed.



(a)

(b)

Figure 6: Input for the segmentation, showing the RGB input image with the mask for the GrabCut initialization laid over it (a), and the segmentation output (b).

Figure 6.b shows the output of the proposed routine, mask refined by GrabCut, portrayed as a binary image: the pixels belonging to the sure foreground (FG) and probable foreground (PR_FG) are depicted black, as opposed to the probable background (PR_BG) and sure background (BG) pixels rendered in white.

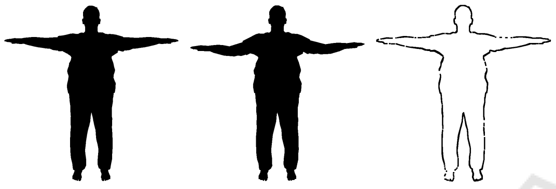
Figure 6b shows the output of the proposed routine—mask refined by GrabCut: the pixels belonging to the sure foreground (FG) and probable foreground (PR_FG) are kept intact, as opposed to the probable background (PR_BG) and sure background (BG) pixels rendered in black.

By comparing Figure 6.a and Figure 6.b, one can see that, although initialized with a seed already quite close to the optimum, the GrabCut algorithm separated the probable foreground from the (probable) background very accurately. Parts of the head and feet were initially estimated as probable background, yet the optimization algorithm successfully brought them back into the (probable) foreground. Similarly, the redundant area around the hands which had been there due to the conservative initialization was removed.

Our approach is further validated on a subset of UPi-S1h images originating from the Leeds Sports Pose dataset, its extended version and the MPII Human Pose Dataset (Lassner et al., 2017). Images on which OpenPose failed to detect human keypoints are not considered for validation as the proposed segmentation approach entirely relies upon them. Of 999 images in the selected subset, 962 successfully passed the routine. To quantitatively express the obtained segmentation results, we compute the Jaccard index (JI), a similarity measure also known as Intersection over Union. The resulting median value over the 962 images equals 0.926.

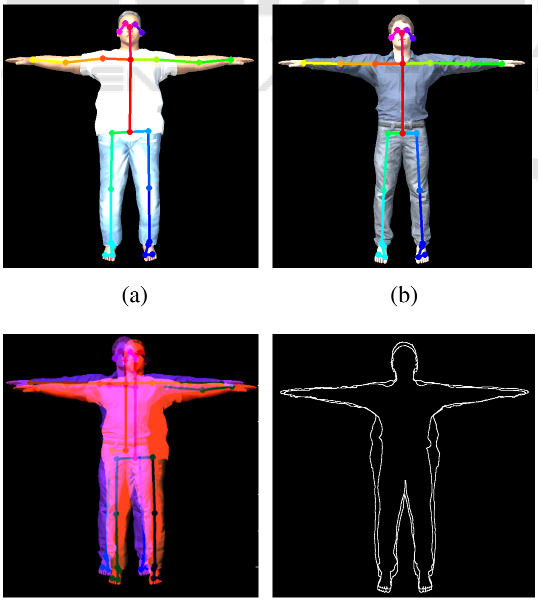
Table 1: Examples for the evaluation of the alignment process showing the randomized skeleton joint offsets used for the transformation. After transforming back to the original pose the Dice Coefficient was calculated using a difference image between the original image and the result of the back transformation. This coefficient shows a correspondence of about 97% and is listed below.

#	Skeleton Joint px-Offset (y-Axis)				Dice Coeff. (%)
	Left Elbow	Left Wrist	Right Elbow	Right Wrist	
0	27	19	33	30	96.9971
1	45	43	48	23	96.3649
2	14	17	24	25	97.2731
3	42	19	20	10	96.1066



(a) (b) (c)

Figure 7: Sample images for the evaluation of the mesh alignment showing the segmentation of the original CGI model (a), the randomly transformed model (b) and the difference image between the original and the back transformed model (c). The difference image was dilated $r = 1$ for a better readability.



(a) (b)

(c) (d)

Figure 8: After alignment, the input person (a) is registered w.r.t. CGI reference silhouette (b) in a manual way utilizing visual inspection (c) and fully-automated (d).

Table 2: Manual and automated registration error calculated from ground truth at varying orientation $\theta = n \cdot 45, n = 0 \ldots 7$ and distance error $d_{err} = |T_x T_{xy}|$.

	manual			auto		
θ	ΔT_x	ΔT_y	d_{err}	ΔT_x	ΔT_y	d_{err}
0	4.18	6.01	7.32	1.68	2.25	2.81
45	0.26	9.96	11.96	0.07	2.91	2.91
90	9.80	12.96	15.46	0.43	0.43	0.61
135	2.22	0.37	2.25	2.28	0.18	2.28
180	1.81	0.34	1.84	1.71	0.80	1.89
225	5.83	3.73	6.91	1.72	0.73	1.87
270	6.50	2.40	6.93	6.44	0.61	6.47
315	0.26	7.49	7.50	2.24	4.33	4.88

5.2 Mesh Alignment

The mesh alignment process is evaluated using multiple binary images of different, segmented CGI 3D models. In addition to the images, also the pose in form of the position of the skeleton joints is known per model. Based on this information, the models are transformed by randomly and independently moving the skeleton joints of the arm by maximal 50 px along the y-axis. Afterwards, the aligned meshes are transformed back using the corresponding original pose. Using the original image and the re-transformed model, a difference image, with a maximal pixel distance of 2 px, and furthermore the Dice Coefficient (DC) (Carass et al., 2020) is calculated to determine the accuracy of the alignment process. The DC results in a deviation of approximately 3%, that only occurs in the model's edge area. This error can be traced back to the contour discretization in the alignment process for improving its performance. The mentioned tests are listed in Table 1 and are shown in Figure 7.

5.3 Registration of Multi-perspective Silhouette Images

The registration utilizing sum of squared errors (SSE) for the contour distance aims at adjusting the position of the current contour w.r.t. the CGI rendered reference shape. Thereby, body shape (slim, obese according to BMI) mismatch between the current person and the CGI and difference in gender do not affect quality of results as seen in the subsequent test cases. Nevertheless, mismatches in skeleton position arising from OpenPose inaccuracies are compensated in a very robust way, cf. Figure 8 and Table 2.

The automated runs take 1153ms on average while visual-guided registration is achieved utilizing MeVisLab software (Ritter et al., 2011). The measured errors for automated registration $\mu_{auto} =$

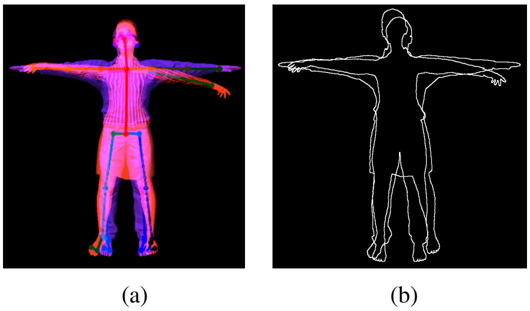


Figure 9: Even for test cases with mismatch in gender and pose (a), the automated contour registration is still feasible applying SSE distance metric (b).

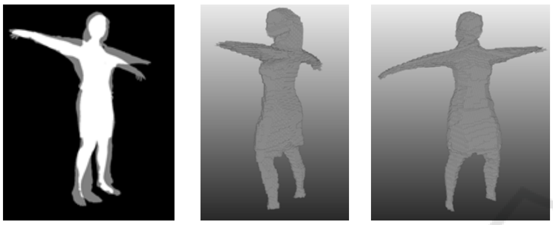


Figure 10: Even if the body pose is not aligned to the reference skeleton, with precise automated registration the subsequent 3D reconstruction is still possible.

2.964, [0.61;6.47] clearly outperform the manual approach $\mu_{auto} = 7.273, [1.84;7.27]$.

Even in case of registering misaligned shapes of a female input dataset with the aligned male reference dataset, the automated registration still leads to robust results with average difference $\varnothing\Delta T_x = 4.54$ and $\varnothing\Delta T_y = 12.53$ compared to manual registration, thus still allowing 3D reconstruction, see Figure 9 and Figure 10.

5.4 Refined Alignment of Human Body Limbs as Kinematic Chains

To test the alignment of the kinematic chains (left/right arm/leg) as final refinement process, target registration tasks are prepared and then used for refinement, see Figure 11. The average error is low with $\varnothing\Delta\beta_1 = 0.034$ and $\varnothing\Delta\beta_2 = 0.074$, see Table 3.

Table 3: Limb alignment performed for right arm. The target transformation for lower/upper arm angles β_1 and β_2 are precisely reached.

	target		result		delta	
#	β_1	β_2	β_1	β_2	$\Delta\beta_1$	$\Delta\beta_2$
1	20.0	-10.0	-20.0	10.0	0.0	0.0
2	11.2	2.3	-11.25	-2.25	0.05	0.05
3	0.0	0.0	0.0	0.0	0.0	0.0
4	-25.0	-17.77	25.0	17.75	0.0	0.02
5	3.87	-7.8	-3.75	7.5	0.12	0.3

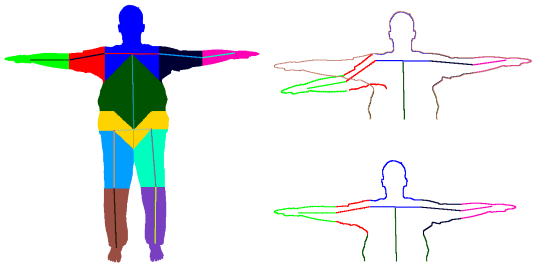


Figure 11: Based on body part segmentation, the limbs are manipulated as kinematic chains until the contours are best matching the reference shape with error evaluated with a distance map.

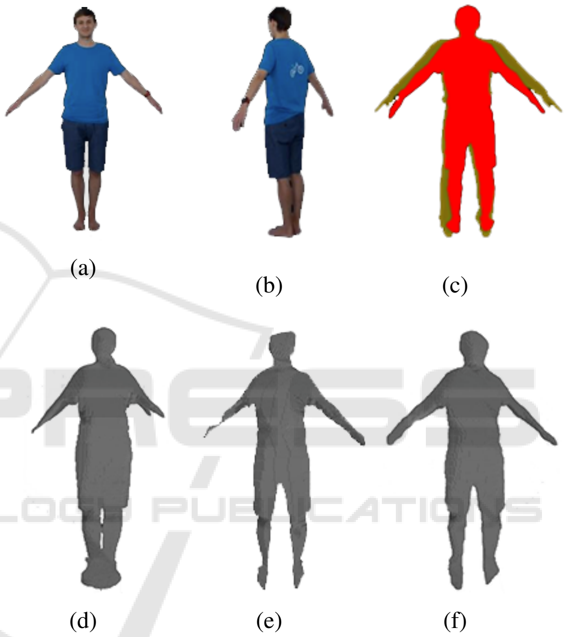


Figure 12: After aligning a real-world human being (a-c), the reconstruction results without alignment (d) are improved (e). In case of applying registration (f), the limb proportions are finally correct.

5.5 3D Silhouette Reconstruction from Aligned and Registered Contours

The proposed alignment steps allow for normalizing the person's silhouette based on the input images of varying orientation and enable 3D reconstruction of a rotating human body from $n = 8$ equidistant views. While rotating on-spot without post-processing is insufficient due to inner body movements, see Figure 12, after alignment and additional registration all body limbs are preserved and well proportioned. In case of thin limbs, the refinement step can lead to additional robustness regarding interaction between alignment and subsequent registration.

Results prove, that registration is the key process in the pre-processing pipeline, see Figure 10 for unaligned but yet auto-registered contours.

6 CONCLUSION AND OUTLOOK

In this research work, a multi-step alignment process is presented to harmonize human body pose acquired during on-spot rotation in A- or T-pose w.r.t. a CGI rendered reference avatar. The shift in position, inner body movements, as well as inaccuracies of OpenPose can be compensated that way allowing for silhouette reconstruction from multiple views. While the ARAP transformation enables the body skeleton's alignment to a reference model keeping the source proportions, the registration process corrects the local position's mismatch and orientation from inaccuracies in OpenPose derived body key-points. OpenPose allows for good quality in orientations and views incorporated in the DL process but lacks precision in unusual orientations. Thus, the body with the skeleton points must be first registered against the CGI rendering silhouette to support reliable anchor positions.

This alignment pipeline is an innovation for moving and elastic objects. As well, this represents an innovation in 3D human reconstruction from monocular camera systems as no depth information is dispensable and no high-level statistical shape model of the human body is required. Instead, a priori knowledge of the human body is kept tight and the specific anatomical variability is preserved.

Currently the CGI renderings can be used in a quite generic way, i.e. male reference template successfully applied for female person or solid level of robustness w.r.t. body mass index. Nevertheless, the subsequent 3D reconstruction can be utilized to derive orientation-specific templates by perspective back-projecting from 3D to 2D after reconstruction of front (0°) and side view (90°) only. This way, the demand for realistic CGI avatars somehow roughly matching the body shape to be reconstructed can be conquered in future leading to more robustness.

A natural progression of this work is to incorporate DL models for object segmentation. This additional knowledge enriches the anthropomorphic model, such as the thickness of limbs, and could improve the GrabCut segmentation algorithm.

Another future development addresses the enrichment of the 3D silhouette reconstruction with depth data either derived from depth sensors of the camera device or DL based estimation of relative depth (Freller et al., 2020) for aesthetic reasons. First results prove that incorporating this additional depth information allows to further shift from the visual hull to the actual body hull. While these improvements do not necessarily affect accuracy of the body sizing measurements, they feature realism of the 3D model visualizations from an immersion point of view. Follow-ups to this work, will focus on 3D reconstruction with aesthetic enrichment and body measurement sizing from reconstructed 3D model.

ACKNOWLEDGEMENTS

This research is initiated by AMB Technology GmbH together with the University of Applied Sciences Upper Austria through project TrueSize S.M.B.S. 21650377 with basic funding program number 872105, sponsored by the Austrian Research Promotion Agency FFG, fed by research budget of the Federal Republic of Austria. Contact: Anna Maria Brunnhofer, CEO.

REFERENCES

Aguilar, W. G., Luna, M. A., Moya, J. F., Abad, V., Parra, H., and Ruiz, H. (2017). Pedestrian detection for uavs using cascade classifiers with meanshift. In *2017 IEEE 11th Int. Conf. on Semantic Computing (ICSC)*.

Bastioni, M., Re, S., and Misra, S. (2008). Ideas and methods for modeling 3d human figures: The principal algorithms used by makehuman and their implementation in a new approach to parametric modeling. In *Proc. of the 1st Bangalore Annual Compute Conf.*

Baumgartner, D., Praschl, C., Zucali, T., and Zwettler, G. A. (2020). Hybrid approach for orientation-estimation of rotating humans in video frames acquired by stationary monocular camera. In *Proc. of the WSCG 2020.*

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools.*

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. on PAMI.*

Carass, A., Roy, S., Gherman, A., Reinhold, J. C., Jesson, A., Arbel, T., Maier, O., Handels, H., Ghafoorian, M., Platel, B., et al. (2020). Evaluating white matter lesion segmentations with refined sørensen-dice analysis. *Scientific Reports*, 10(1):1–19.

Chew, L. P. (1989). Constrained delaunay triangulations. *Algorithmica*, 4(1-4):97–108.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models - their training and application. *Comp. Vision and Image Underst.*, 61(1).

Douglas, D. H. and Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122.

Fang, H.-S., Lu, G., Fang, X., Xie, J., Tai, Y.-W., and Lu, C. (2018). Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *Proceedings of the IEEE CVPR*.

Fit3D (2020). Ein berührungsloses erlebnis mit 3d-körperscans.

Freller, A., Turk, D., and Zwettler, G. A. (2020). Using deep learning for depth estimation and 3d reconstruction of humans. In *Proc. of the 32nd European Modeling and Simulation Symposium, Vienna, Austria*.

Heindl, C., Bauer, H., Ankerl, M., and Pichler, A. (2015). Reconstructme sdk: a c api for real-time 3d scanning. In *Proc. of the 6th Int. Conf. on 3D Body Scanning Technologies, Switzerland*.

Hidalgo, G. and Fujisaka, Y. (2019). Pose output format (body 25).

Jacobson, A., Panozzo, D., et al. (2018). libigl: A simple C++ geometry processing library. https://libigl.github.io/.

Kocabas, M., Athanasiou, N., and Black, M. J. (2019). Vibe: Video inference for human body pose and shape estimation.

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M. J., and Gehler, P. V. (2017). Unite the people: Closing the loop between 3d and 2d human representations.

Li, X., Li, H., Joo, H., Liu, Y., and Sheikh, Y. (2018). Structure from recurrent motion: From rigidity to recurrency.

Lin, H. and Wu, J. (2008). 3d reconstruction by combining shape from silhouette with stereo. In *2008 19th Int. Conf. on Pattern Recognition*, pages 1–4.

Liu, H., Xu, T., and Wang, X. (2013). Related hog features for human detection using cascaded adaboost and svm classifiers. *Lecture Notes in Computer Science book series*, 7733.

Liu, T. and Stathaki, T. (2017). Enhanced pedestrian detection using deep learning based semantic image segmentation. In *2017 22nd Int. Conf. on Digital Signal Processing (DSP)*, pages 1–5.

Marchal, G. and Lygren, T. (2017). The microsoft kinect: validation of a robust and low-cost 3d scanner for biological science.

Mazareanu, E. R. (2020). statista: Return deliveries - costs in u.s. 2017-2020. https://www.statista.com/statistics/871365/reverse-logistics-cost-united-states, last visited 2020-07-29.

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, 2014(239).

Mulayim, A., Yılmaz, U., and Atalay, M. V. (2003). Silhouette-based 3-d model reconstruction from multiple images. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics*, 33:582–91.

Murphy, T. (2020). Improve conversion rates with your returns policy. https://www.hiplee.com/blog/improve-conversion-rates-with-your-returns-policy/, last visited 2020-07-29.

Orendorfff, A. (2019). The plague of ecommerce return rates and how to maintain profitability. https://www.shopify.com/enterprise/ecommerce-returns, last visited 2020-07-29.

Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., and Schiele, B. (2017). Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67.

Pointner, A., Krauss, O., Freilinger, G., Strieder, D., and Zwettler, G. A. (2018). Model-based image processing approaches for automated person identification and authentication in online banking. In *Proc. of the EMSS2018*.

Ritter, F., Boskamp, T., Homeyer, A., Laue, H., Schwier, M., Link, F., and Peitgen, H. . (2011). Medical image analysis. *IEEE Pulse*, 2(6):60–70.

Rother, C., Kolmogorov, V., and Blake, A. (2004). " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314.

Seitz, S. M. and Dyer, C. R. (1999). Photorealistic scene reconstruction by voxel coloring. *Int. Journal of Computer Vision*, 35:151–173.

Sizestream (2020). Best-in-class accuracy and speed in a commercial 3d body scanner customizable platform.

Song, Z., Yu, J., Zhou, C., Tao, D., and Xie, Y. (2013). Skeleton correspondence construction and its applications in animation style reusing. *Neurocomputing*, 120:461 – 468.

Sorkine, O. and Alexa, M. (2007). As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116.

Statista (2012). Share of online orders that were returned in 2012 (by product category). [Online; accessed August 9, 2020].

Suzuki, S. et al. (1985). Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. (2017). Learning from synthetic humans. *CoRR*, abs/1701.01370.

Wei, G., Lan, C., Zeng, W., and Chen, Z. (2019). View invariant 3d human pose estimation. *CoRR*.

Xiang, Y., Schmidt, T., Narayanan, V., and Fox, D. (2017). Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *CoRR*, abs/1711.00199.

Yu, Y., Makihara, Y., and Yagi, Y. (2019). Pedestrian segmentation based on a spatio-temporally consistent graph-cut with optimal transport. *IPSJ Transactions on Computer Vision and Applications*, 11.

Zatsiorsky, V. (1998). Kinematics of human motion. *American Journal of Human Biology*, 10.

Zeng, D., Chen, X., Zhu, M., Goesele, M., and Kuijper, A. (2018). Background subtraction with real-time semantic segmentation.