TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell

Hayato Onizuka  
Kyushu University  
mound028810@gmail.com

Zehra Hayırcı  
Technical University of Munich  
zebrahimayirci@gmail.com

Diego Thomas  
Kyushu University  
thomas@ait.kyushu-u.ac.jp

Akihiro Sugimoto  
National Institute of Informatics  
sugimoto@nii.ac.jp

Hideaki Uchiyama  
Kyushu University  
uchiyama@limu.ait.kyushu-u.ac.jp

Rin-ichiro Taniguchi  
Kyushu University  
rin@kyudai.jp

Figure 1: We present a tetrahedral volumetric representation of the human body and a method called TetraTSDF that is able to retrieve the detailed 3D shape of a person wearing loose clothes from a single 2D image.

Abstract

Recovering the 3D shape of a person from its 2D appearance is ill-posed due to ambiguities. Nevertheless, with the help of convolutional neural networks (CNN) and prior knowledge on the 3D human body, it is possible to overcome such ambiguities to recover detailed 3D shapes of human bodies from single images. Current solutions, however, fail to reconstruct all the details of a person wearing loose clothes. This is because of either (a) huge memory requirement that cannot be maintained even on modern GPUs or (b) the compact 3D representation that cannot encode all the details. In this paper, we propose the tetrahedral outer shell volumetric truncated signed distance function (TetraTSDF) model for the human body, and its corresponding part connection network (PCN) for 3D human body shape regression. Our proposed model is compact, dense, accurate, and yet well suited for CNN-based regression task. Our proposed PCN allows us to learn the distribution of the TSDF in the tetrahedral volume from a single image in an end-to-end manner. Results show that our proposed method allows to reconstruct detailed shapes of humans wearing loose clothes from single RGB images.

1. Introduction

Detailed 3D shapes of the human body reveal personal characteristics that cannot be captured with standard 2D pictures. Such information is crucial for many applications in the entertainment industry (3D video), business (virtual try-on) or medical use (self-awareness or rehabilitation). The first systems that built 3D models of the human body were designed to work in controlled settings using laser scanners, multi-view calibrated camera arrays or markers. These systems were hard to set up, not affordable, and offering limited application areas.

In the last decade, consumer-grade depth cameras have been successfully used to build 3D models of the human body [45]. However, high quality depth cameras are still not available to a large part of the consumers (most of smartphones are still not equipped with depth cameras). Moreover, consumer-grade depth cameras do not work well in outdoor environments. As a consequence, methods that can efficiently reconstruct detailed 3D human shapes in unconstrained environments are needed. In this work, we focus on the task of detailed 3D human body reconstruction in a single shot with a standard RGB camera.

In the literature, there are two strategies to generate a 3D...
CNN-based methods are promising for reconstructing 3D human bodies from a single color image because they have the potential to capture detailed and complex features (such as clothes wrinkles). However, several limitations exist. The main problem stems from the volumetric TSDF representation which is used to regress the 3D shape. The resolution of the volumetric representation has to meet the fine details of the human body, meaning that a considerable amount of memory is required. However, such memory constraint is hard to be maintained even on modern GPUs.

The key challenge for achieving higher accuracy in 3D human body shape reconstruction is to define a more compact 3D human body representation in memory that still allows casting the problem as a well-adapted regression task. One solution may be displacement mapping [1]. However, this compact representation inevitably loses some details in shape because its dependency to the SMPL mesh. For instance, occluded parts in non-convex areas or garments like shoes or gloves cannot be reconstructed by using displacement mapping.

In this paper, we propose a new volumetric 3D body representation for end-to-end 3D body shape regression from a single color image. Our proposed 3D body representation is based on a tetrahedral TSDF field embedded into a human-specific outer shell. The outer-shell is built from a coarse version of the SMPL model [19] and can be fitted to a human body using the SMPL pose and shape parameters. The tetrahedral TSDF field is built at the summits of a tetrahedral volumetric grid defined by the outer shell. We also propose a new network to estimate the tetrahedral TSDF field from a single color image that combines CNN and our proposed Part Connection Network (PCN).

Our contributions are threefold: (1) a new 3D body volumetric representation that is compact, dense, accurate and yet well suited for CNN-based regression tasks; (2) a method to generate high quality TSDF fields from ground truth (GT) 3D human body scans; and (3) a new CNN-PCN based hourglass network for end-to-end regression of 3D human body shape from a single color image.

2. Related work

Fitting a parametric 3D model to the input 2D color image has been the standard way to reconstruct a 3D shape from a single 2D image for a long time. Recently, using CNNs has proven to be a powerful alternative. Here, we review related works that use both of these strategies, with a particular focus on the human body shape reconstruction problem.

2.1. Template model fitting

The classic approach to estimate a 3D shape of an object from a single color image is to fit a template 3D model so that it matches its 2D projection while satisfying some constraints (e.g., [4, 5, 19]). Landmark-guided non-rigid registration of 3D templates to 2D or 2.5D inputs have been widely studied. Lu et al. [21], for example proposed to use facial landmarks to fit a deformable face model to 2.5D data. Cashman et al. [5], proposed to represent a 3D deformable model with a linear combination of subdivision surfaces. Such model can be fitted to a collection of 2D images by manually providing some key-points and the silhouette of the object to be reconstructed.

Recovering the 3D human shape from a single RGB image is an open-problem in the field. There are only few proposed techniques that deal with complex poses and deformations. Most of the techniques rely either on a pre-scanned model of the subject [8, 13, 18, 32, 38] or a template model [2, 3]. In [13], the authors propose to first scan the 3D model of a person using a multi-view reconstruction system. Then, the reconstructed 3D model is non-rigidly aligned to the RGB video in real-time. Similarly, in [12], Bogo et al. [4], propose to optimize the parameters of a parametric template model given input images and poses by using many cues like silhouette overlap, height constraint and smooth shading. The authors use the SMPL model [19] to recover various parameters such as pose and shape from a single RGB image. Recently, Kolotouros et al. [17] proposed a method to estimate SMPL parameters by developing a self-improving loop combining CNN and optimization method. However, template-based methods fail to capture loose clothing and thus only reconstruct bare human body.

2.2. Convolutional Neural Network regression

Recently, CNNs have brought new possibilities to many domains in computer vision. 3D shape reconstruction from a single image is one of those areas that strongly developed with the availability of new CNN tools.

Inspired by the extraordinary performance of CNNs for segmentation tasks, several methods have been proposed that represent the 3D shapes as binary occupancy maps [6, 11, 35, 40, 41]. If the task of estimating the 3D surface is expressed as a segmentation problem, then CNNs
can predict outside and inside voxels. For example, Wu et al. [42] proposed an extension of 2D CNNs for the case of volumetric outputs. Further optimization improvements were proposed in [44] and [46]. In the case of 3D face regression, Jackson et al., [15] proposed a method for direct regression of a volumetric representation of the face using CNN. By using probabilistic assignment, smooth surfaces could be obtained. Varol et al. [36], extended this method to full body shape regression. All these methods share the common limitation that the memory consumption scales cubically with the shape resolution. Even with modern GPUs, volumetric regression networks only work with low resolution grid. Then, only coarse 3D models can be generated.

Riegler et al. [28], proposed to use octrees to reduce the memory usage and adapt the CNNs to predict high-resolution occupancy maps providing the tree structure is known in advance. However, the method can not be applied to reconstruct 3D human bodies with different poses because the tree structure changes with every new input. In [34], Tararchenko et al. proposed a technique to overcome this problem by also predicting the tree structure. However, training the network to learn the sparse structure of octrees is effortful. Recently, Saito et al. [30] proposed a memory-efficient method by handling each 3D point individually. They reported high accuracy 3D reconstruction results on a private dataset while using reasonable amount of memory. However, this method jointly estimates the 3D shape and body pose while 3D body pose estimators have made significant progress recently and achieved high accuracy results (e.g., [14] reported an average error of less than 2 cm for the 3D joint position). We reason that the tasks should be separated: one task for body pose estimation and one task dedicated to the 3D shape estimation.

Recently, Alldieck et al. [1] have proposed to use displacement mapping on top of a template human model to represent the 3D human body with loose clothes. This not only fits well to the CNN formulation but also it requires low memory. However, it has a severe limitation that it can not reconstruct convex parts like inside cloth wrinkles. In addition, because the template human model is naked with fingers in hands and feet and deviation mapping can only encode displacement in the normal vector direction, reconstructing shoes or gloves for example is not possible. In the meantime Gabeur et al. [10] proposed a method that predicts depth maps from visible and hidden side by using GAN to "mould" the 3D human body.

We reason that when reconstructing the human body with loose clothes, the implicit volumetric TSDF is the best representation to handle various shapes. We observe that in the regular grid, many of the voxels in the 3D bounding box around the person are actually unnecessary. Therefore, we propose a new tetrahedral 3D shape representation that is able to reduce the memory consumption drastically while allowing to reconstruct high resolution 3D shapes.

### 3. Proposed tetrahedral representation

The Truncated Signed Distance Function was first introduced by Curless and Levoy in [7] to represent 3D surfaces and has been extensively used in modern RGB-D simultaneous localisation and mapping (SLAM) systems [23, 24, 27, 29]. At any point in the 3D space, the TSDF function takes the signed distance to the 3D surface as value. These values are truncated between $-1$ and 1 for practical implementation reasons. In general the TSDF is sampled in a regular grid of (rectangular) voxels and the 3D mesh of the surface can be extracted using well established algorithms such as the Marching Cubes [20] or ray tracing.

Our objective here is to reduce the irrelevant field around the human body while covering the meaningful space, which we call the outer shell. We propose to modify the well known SMPL model [19] to create this outer shell. The SMPL model has well defined pose and shape parameters that can be fitted to any 3D human dataset, skeleton or even RGB image (using CNNs for example).

#### 3.1. Coarse human outer shell

We propose to inflate the SMPL neutral body model so that once fitted to the input it covers the entire body as well as the loose clothes. We also propose to remove shape details of the SMPL model (such as nose, mouth etc ...). Our pipeline is illustrated in Figure. 2. We reason that we do not need details at the surface of the outer-shell because the details will be encoded into the TSDF field. Therefore we...
first create a ground truth dataset to supervise the learning process. In our case, the training dataset consists of a set of pairs of one 2D image and one corresponding dense tetrahedral TSDF field (i.e., the set of TSDF values for each voxel summit). To build a large amount of dense TSDF fields from a set of publicly available ground truth human 3D scans we need an efficient algorithm. In this section we detail our proposed algorithm to generate such ground truth training dataset.

Firstly, we fit the coarse model to the GT 3D scan by optimizing the SMPL pose and shape parameters given the GT 3D skeleton. Then we compute the TSDF value for each voxel of the outer shell. Here arise two problems: (1) the GT 3D mesh is sparse, so the standard signed distance between a voxel summit and its closest point is significantly different than the signed distance to the surface. (2) In some poses parts of the coarse model may overlap, which may result in ghost effects (for example, a part of the torso may be encoded into the voxels that correspond to the arm).

To solve the first problem we compute the TSDF values as the point-to-plane distance between the voxel summit and the tangent plane of the closest point. This is a reasonable approximation of the point-to-surface signed distance (and it can be computed quickly) when the voxel is close to the surface. However, the approximation may become completely wrong when the voxel is far from the surface. To overcome this, we truncate the TSDF values based on a threshold on the euclidean point-to-point distance between the voxel summit and its closest point.

\[
TSDF(v) = \begin{cases} 
\frac{\hat{n} \cdot (v - \hat{v})}{\sigma(\hat{n} \cdot (v - \hat{v}))} & \text{if } ||v - \hat{v}||_2 \leq \tau \\
\frac{1}{\sigma(1)} & \text{if } ||v - \hat{v}||_2 > \tau
\end{cases},
\]

where \(v\) is a voxel summit, \(\hat{v}\) is the closest point of \(v\) in the GT 3D mesh, and \(\hat{n}\) is the normal vector at \(\hat{v}\). The threshold \(\tau\) is set to 3 cm in the experiments and the function \(\sigma(\cdot)\) returns the sign of its argument.

In order to solve the second problem we compute the TSDF values from the 3D scan warped into a star pose, where the body parts of the outer shell do not penetrate each other (see fig.4). To this end we first connect each vertex of the GT 3D mesh to its corresponding skeletal nodes with appropriate weight. Then we warp the 3D vertex coordinates to the T-shape by blending the 3D transformation of the attached skeletal joints. Once all vertices have been warped, we compute the dense TSDF field of the outer-shell as explained above.

4. Detailed 3D shape regression

We propose an end-to-end regression network to estimate the TSDF values in the tetrahedral volume from a single image. Our proposed network takes as input a single 2D

---

1. We used the Blender software (https://www.blender.org) to do the dilation, down-sampling and up-sampling processes.
Figure 5: Our proposed network is a combination of CNN and PCN. Given a 3D pose data, it allows us to regress the 3D shape of a person wearing clothes in an end-to-end fashion.

image and outputs the dense TSDF field. From this TSDF field a detailed 3D surface of the human can be extracted by using the Marching Tetrahedra algorithm. The estimated 3D mesh can be re-posed using SMPL pose parameters obtained by any state-of-the-art 3D pose estimation method.

The standard approach to regress TSDF in volumetric data is to use stacked hourglass networks [25]. In this way, the input image is encoded into a feature vector, which is then decoded into the volumetric grid. To this end, CNNs are used to build the network with the help of the well-organized data into uniform grids of pixels and voxels. In our case the volumetric data is embedded into the tetrahedral mesh, which does not have the uniform grid organization. As a consequence, state-of-the-art CNN hourglass architecture cannot be used directly.

We propose a new hourglass structure to regress the TSDF value of each vertex in the tetrahedral voxel model. Our proposed network combines CNN to encode the input image and a new Part Connection Network (PCN) to decode the feature vector (Figure 5 illustrates our proposed network). The originality of our network resides in the later part. We propose to build several tetrahedral layers by down-sampling the full resolution outer shell volumetric model (see fig. 5), and then we propagate the features in the upper layer to output the TSDF field. Note that the number of voxels in the tetrahedral model is too large to directly use a fully connected network (the number of parameters near the final layer consumes about 90GB of memory in the case of a fully connected network). Instead we propose a Part Connected Network with partially connected layers, where connections between successive layers are done only in-between the same body parts (which consumes only 0.025 GB of memory).

For each layer $l$ we define the partial connections using the following sparse adjacency matrices.

$$A^{(1)^{l}} = \begin{bmatrix} A^{(1)^{l}} \\ A^{(2)^{l}} \\ \vdots \\ A^{(n_{out})^{l}} \end{bmatrix}.$$ (2)

$$A^{(n)^{l}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_{in}} \\ a_{21} & a_{22} & \cdots & a_{2n_{in}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn_{in}} \end{bmatrix},$$ (3)

where $a_{ij} = \begin{cases} 1 & \text{if } j = adj(n)^{[i]} \\ 0 & \text{otherwise.} \end{cases}$

$$A^{2} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}.$$ (4)
where \( n_{\text{in}}, n_{\text{out}} \) is the number of input and output nodes, \( m \) is the number of adjacent nodes for the \( n \)-th output node, and \( \text{adj}(n) \) is the list of indices of input nodes that are connected to the \( n \)-th output node (and \( \text{adj}(n)[i] \) is the \( i \)-th element in the list). There is no guarantee that all of the output nodes have the same number of adjacent nodes, so we flatten the input features by using the matrix \( A^1_l \), and reshape it to the shape of the output features by using \( A^2_l \).

With these matrices, the features from two successive layers \((l, l+1)\) are transmitted as follows.

\[
f_{l+1} = \sigma(A^2_l(A^1_l f_l \circ W_l)), \tag{5}
\]

where \( f_l, f_{l+1} \) are the input and output features of the layer and \( W_l \) is the variable weight matrix for all edges. Also, \( \circ \) denotes the element-wise product of two matrices and \( \sigma \) denotes the activation function.

To define the adjacency matrices \( A^1_l \) and \( A^2_l \) we need to identify the adjacency lists \( L \) that connect the input nodes and the output nodes of all layers. To create these adjacency lists, we focus on the locality of the nodes in the tetrahedral model. Concretely, each node in the \( l \)-th layer is connected to the \( k \)-nearest neighbors in the \( l+1 \)-th layer (we used \( k = 5 \) in our experiments). We reason that the tetrahedral voxel model has a human body shaped graph structure and that we can consider that distant nodes have a weak relationship with each other (like between toes and fingertips). As a consequence, by connecting only the near nodes to the next layer and not connecting the distant nodes, the number of learning parameters is drastically reduced without loosing much information from the features from the previous layer.

We design the network structure so as not to lose features while reducing the number of variable parameters by connecting only the adjacent nodes. However, when connecting the last CNN layer of the first half of the network to the first PCN layer of the latter half, there is no notion of adjacency because nodes from the CNN layer do not have the shape of the human body. Therefore, we use a fully connected layer between the CNN and the PCN networks.

\section{5. Experiments}

We evaluate qualitatively and quantitatively the ability of our proposed method to reconstruct 3D shapes of the human body from a single image by using publicly available datasets.

In all our experiments, we used a volumetric resolution (i.e., average distance between adjacent summits in the tetrahedral volume) of about 1 cm (which corresponds to about \( 2.6 \times 10^5 \) voxel summits). We evaluated our network on SURREAL \cite{37} and Articulated \cite{38} datasets. In the evaluation using SURREAL, we strictly followed the protocol as explained by the authors in \cite{37}. Our network was tested on subjects not appearing in the training dataset.

We compared our proposed method with other recent works that have made their code publicly available (\cite{1,26,36}). Note that for Tex2Shape \cite{1} only the code for testing is available and so we built a discriminator network for the training of Tex2Shape referring to their paper.

\begin{table}[ht]
\centering
\begin{tabular}{|c|c|c|}
\hline
 & SURREAL & Articulated \\
\hline
SMPLify-x & n.a. & 9.61 \\
BodyNet & 6.38 & 7.22 \\
Tex2Shape & n.a. & 0.72 \\
Ours & \textbf{5.14} & \textbf{0.43} \\
\hline
\end{tabular}
\caption{Quantitative comparison of results obtained with our method, BodyNet and Tex2Shape on the SURREAL \cite{37} (naked) and Articulated \cite{38} (clothed) datasets.}
\end{table}

This allows us to test the generalizability of our proposed method. In Articulated, we trained our network on 80\% of the data and tested on the remaining 20\%. Our method was tested on poses (and thus deformations) not appearing in the training dataset. This allows us to test the ability of our proposed methods to reconstruct detailed 3D shapes such as clothes wrinkles. In the training, we employed the mean square error loss function in the last layer and \text{reLU} as an activation function in the hidden layers. It took about 3-4 hours to train the network for Articulated dataset, with using a batch size of 5 and a single GTX 1080 GPU.

We compared our proposed method with other recent works that have made their code publicly available (\cite{1,26,36}). Note that for Tex2Shape \cite{1} only the code for testing is available and so we built a discriminator network for the training of Tex2Shape referring to their paper.

\subsection{5.1. Comparative evaluation on SURREAL dataset}

We evaluate qualitatively and quantitatively the ability of our proposed method to reconstruct 3D shapes of the human body from a single image by using publicly available datasets.

In all our experiments, we used a volumetric resolution (i.e., average distance between adjacent summits in the tetrahedral volume) of about 1 cm (which corresponds to about \( 2.6 \times 10^5 \) voxel summits). We evaluated our network on SURREAL \cite{37} and Articulated \cite{38} datasets. In the evaluation using SURREAL, we strictly followed the protocol as explained by the authors in \cite{37}. Our network was tested on subjects not appearing in the training dataset.

\begin{figure}[ht]
\centering
\includegraphics[width=0.5\textwidth]{figure6.png}
\caption{An example of a 3D mesh obtained with our method and with BodyNet on the SURREAL dataset.}
\end{figure}
Figure 7: Comparative reconstruction results obtained with our proposed method and other related works. From left to right: input image, ground truth, our method, Tex2Shape [1], SMPLify-X [26], BodyNet [36].

Table 1 shows the quantitative comparison. For the metric we used the Chamfer distance between the ground truth 3D scans and the reconstructed 3D meshes. As we can see from these results our proposed method was able to reconstruct accurate dense 3D shapes of the human body from only one single image. We obtained better results that BodyNet because our proposed tetrahedral representation allows for reconstruction at higher resolution.

5.2. Comparative evaluation on Articulated dataset

To confirm the advantages of our proposed method for reconstructing detailed 3D shapes of humans with loose clothes from a single image we compared our method with the most recent state-of-the-art method Tex2Shape [1]. For qualitative and quantitative evaluation, we used the publicly available dataset called Articulated [38] that contains sequences of human wearing loose clothes in many poses with ground truth 3D scans.

Unfortunately, the dataset used in [1] is not publicly available so we could not directly compare our proposed method with [1] on their own dataset. We built a discriminator network for the training of Tex2Shape referring to their paper [1] and trained their network on ARTICULATED dataset using the exact same train/test split as used with our proposed method. The comparative results shown in Figure 7 and Table 1 shows advantages of our proposed method compared with Tex2Shape. Note that as in [1] we show the results re-posed with the ground truth 3D pose. However, any third party 3D pose estimation could also be used to estimate the 3D pose from the input image ([22, 39]). Figure
Figure 8: Visualization of the Euclidean distances between the vertices of reconstructed 3D mesh and the closest points in the GT scan. These errors are represented with heatmaps and mapped over the reconstructed 3D mesh for each method.

7 shows some representative results from the test dataset. We also compared the results obtained with our method with those obtained with other closely related works. As seen in figures 7 and 8, our proposed method outperforms all other previous works. Figure 8 shows the heat maps of errors of the 3D reconstructed models obtained by our method and other related works. As we can see our proposed network could successfully recover the details of the loose clothes, even in occluded areas. As we can see in Table 1 we observed an average error around 0.5 cm in the 3D meshes reconstructed with our proposed method.

As we can see in the circled areas of figure 7, the strongest and clearer advantage of our proposed method over the related works is that our proposed method can retrieve both hand details such as fingers in the correct position (no details can be obtained with BodyNet) or in the inverse reconstruct the shoes around the feet (fingers still remain with Tex2Shape, giving an unpleasant effect for people wearing shoes). Moreover, with our proposed method we do not need to estimate finger poses. As we can see in Figure 7 although the hand pose is unknown (this is why the hand is always open in the results obtained with Tex2Shape) our proposed method could successfully reconstruct the detailed hand in the correct pose.

5.3. Network analysis

We analysed the performance of our network by changing the parameters. Figure 9 (a) shows the convergence speed of our network during training. Figure 9 (b) shows the performance of our network when changing the number of connections in the PCN. As expected, there is a trade-off between the number of connections (i.e., memory consumption) and accuracy.

6. Conclusion

We proposed a method for reconstructing fine detailed 3D shapes of human body wearing loose clothes from a single image. We introduced a novel 3D human body representation based on a tetrahedral TSDF field embedded into a coarse human outer shell. We also designed CNN-PCN network to regress the tetrahedral TSDF field in an end-to-end fashion. Our qualitative and quantitative comparative experiments using public datasets confirmed the ability of our proposed method to reconstruct detailed shapes with loose clothes. The results demonstrated out-performance of our method against the current state-of-the-art on these datasets. Several possible improvements are left for future work, such as combining our PCN with the Graph Convolutional Network [9][43].

Acknowledgement

A part of this work was supported by JSPS KAKENHI Grant Number JP19K20297 and Microsoft CORE 16 Grant.
References


[27] Victor Adrian Prisacariu, Olaf Kähler, Stuart Golodetz, Michael Sapienza, Tommaso Cavallari, Philip HS Torr, and


