Template based Human Pose and Shape Estimation from a Single RGB-D Image

Zhongguo Li, Anders Heyden and Magnus Oskarsson
Centre for Mathematical Sciences, Lund University, Sweden

Keywords: Human Body Reconstruction, SMPL Model, 2D and 3D Pose, Pose and Shape Estimation.

Abstract: Estimating the 3D model of the human body is needed for many applications. However, this is a challenging problem since the human body inherently has a high complexity due to self-occlusions and articulation. We present a method to reconstruct the 3D human body model from a single RGB-D image. 2D joint points are firstly predicted by a CNN-based model called convolutional pose machine, and the 3D joint points are calculated using the depth image. Then, we propose to utilize both 2D and 3D joint points, which provide more information, to fit a parametric body model (SMPL). This is implemented through minimizing an objective function, which measures the difference of the joint points between the observed model and the parametric model. The pose and shape parameters of the body are obtained through optimization and the final 3D model is estimated. The experiments on synthetic data and real data demonstrate that our method can estimate the 3D human body model correctly.

1 INTRODUCTION

Human body reconstruction is needed in many applications such as virtual and augmented reality, video games and medical research. Various approaches have been proposed to obtain 3D models of human bodies during past decades. However, it is still a challenging problem from using one single image since human body may have high complexity such articulation, self-occlusion, clothing and so on. Commercial laser scanning system can acquire 3D models with high accuracy. But it is not acceptable for ordinary consumers because of high cost and cumbersome equipment. Some other methods based on RGB-D image sequences are also popular to build the 3D model (Izadi et al., 2011; Innmann et al., 2016; Newcombe et al., 2015; Slavcheva et al., 2017). They often requires the object to be static and rigid during acquiring the image sequence, which is not appropriate for a moving person.

It is possible now to create 3D human body models through combining human pose estimation and some parametric models, especially with the implementation of deep neural network to 2D- (Xu et al., 2017) and 3D- (Mehta et al., 2017) human pose estimation. These methods (Bogo et al., 2015; Huang et al., 2017; Kanazawa et al., 2018) utilize a parametric body model to fit 2D or 3D human pose which can be well estimated by deep neural networks. They are more efficient and convenient because they do not need to use expensive devices and too much images with different views. This makes it possible to reconstruct 3D human body through only one image. The key step of these methods is to construct an objective function to measure the difference between the parametric model and the person in the image. In many cases only 2D pose is utilized to construct the objective function, which can not fully represent the joints points of the human body in 3D space. Currently, many methods based on deep neural networks have been proposed to estimate 3D human pose, but are not in general as accurate as 2D pose estimation.

In this paper we propose to use both 2D human pose which is estimated by deep neural network and 3D human pose which is calculated from a depth image to construct objective function. The 3D human body models based on the skinned multi-person linear model (SMPL)(Loper and Black, 2014) are then estimated by optimizing the objective function. This parametric model gives means to better to describe the pose and shape of the human body. More specifically, the 2D human pose is estimated by Convolutional Pose Machines (CPM) (Wei et al., 2016) which predicts the 2D joint points by stacking convolutional neural networks. For the 3D pose, we use a Kinect sensor to acquire RGB-D images and the depth im-

Li, Z., Heyden, A. and Oskarsson, M.
Template based Human Pose and Shape Estimation from a Single RGB-D Image.
DOI: 10.5220/0007383605740581
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 574-581
ISBN: 978-989-758-351-3

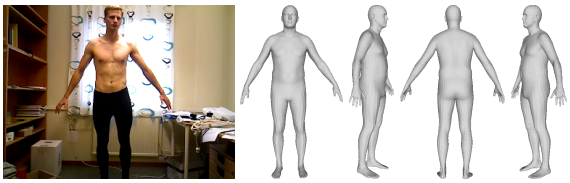


Figure 1: The original image and the corresponding 3D model from different view reconstructed by our method.

age provides depth information which can be used for the computation of the corresponding 3D joint points. Then, an objective function considering the difference between the estimated pose and the pose of the SMPL model is defined. Through minimizing the objective function, pose and shape parameters of the SMPL model are estimated, which creates a 3D human body model with certain accuracy. We compare our method with an existing method which only considers 2D pose to reconstruct the 3D human body model and the experiments quantitatively and qualitatively show that our method has better performance both on the accuracy of pose parameters and 3D vertices.

In summary the contribution of our method is that we use both 2D and 3D pose information based on the SMPL human body model to estimate the 3D model from a single RGB-D image.

2 RELATED WORK

Non-parametric model based methods often reconstruct 3D scenes directly from images acquired by a camera or depth sensor from multiple views. Registration is an important step for these methods, which has been solved well for rigid transformations. However, the human body is inherently non-rigid and it is a great challenge for non rigid registration. KinectFusion (Izadi et al., 2011) is a well-known algorithm to create 3D models in real time by incrementally merging the partial scans from a moving RGB-D sensor. Inspired by this research, a number of methods (Ari et al., 2014; Cui et al., 2012) have been proposed to reconstruct 3D human body models for static persons. However, they are not ideal approaches for the reconstruction of the dynamic human body due to its complex pose and the non-rigidity of the human body. In order to tackle the non rigid problem, DynamicFusion (Newcombe et al., 2015) is the pioneering work and can reconstruct scene geometry in real time for a slowly moving person. Some other methods such as VolumeDeform (Innmann et al., 2016), KillingFusion (Slavcheva et al., 2017) and BodyFusion (Yu et al., 2017) have been proposed to improve the results based on the DynamicFusion. However, these approaches only show good performance when the motion of the human body is slow and only half of the body is reconstructed. For more complex conditions, researchers try to use multiple Kinect sensors or several calibrated cameras to create the 3D human body models, which can acquire accurate 3D models. In (Dou et al., 2016), the authors propose to use eight Kinects to create accurate 3D models for dynamic scenes. Multiple cameras are also used in (Leroy et al., 2017) used to obtain the 3D shape of the human body with complex actions. Although it yields more accurate 3D human body models by using multiple devices, it is inconvenient and expensive to require such an environment.

Parametric model-based methods reconstruct 3D body models through using a template which always encodes the pose and shape parameters to fit the observed images. Some methods firstly scan a model as the template, and then insert a skeleton into the template and use the template to fit other dynamic data. In (Guo et al., 2015; Li et al., 2009) the authors propose novel nonrigid registration algorithms and they implement the algorithms to register the prescanned models to the partial scans of Kinect. (Zhang et al., 2014) utilize KinectFusion to get the template of the human body before fitting the input data. Some other algorithms based on this pipeline are also presented in (Xu et al., 2017; Zollhofer et al., 2014). However, prescanning a templete depends on the algorithm of reconstructing algorithm for rigid objects. In order to better represent the human body, a number of statistic human body models based on training on thousands of individuals have been developed, such as SCAPE (Anguelov et al., 2005), SMPL (Loper et al., 2015) and so on (Pons-Moll et al., 2015). The SCAPE model is used in (Weiss et al., 2011) to fit the depth image acquired by Kinect. Authors in (Bogo et al., 2015) propose a detailed body reconstruction method using an improved SCAPE model called Delta. There are also many methods in which the SMPL model is used. In (Bogo et al., 2014), the authors propose to use a CNN model to extract 2D joint points and used the information to fit a parametric model. Huang et al (Huang et al., 2017) use a similar idea but they extend it to multiple view problems. In (Kanazawa et al., 2018) an end-to-end adversarial learning method is used to estimate the human pose and shape by fitting a parametric model. Alldieck et al (Alldieck et al., 2018) propose to get the 3D human body model from video based on the SMPL model using pose information and transforming silhouettes to visual hull. In this paper we also use a template based method to reconstruct the 3D body model.

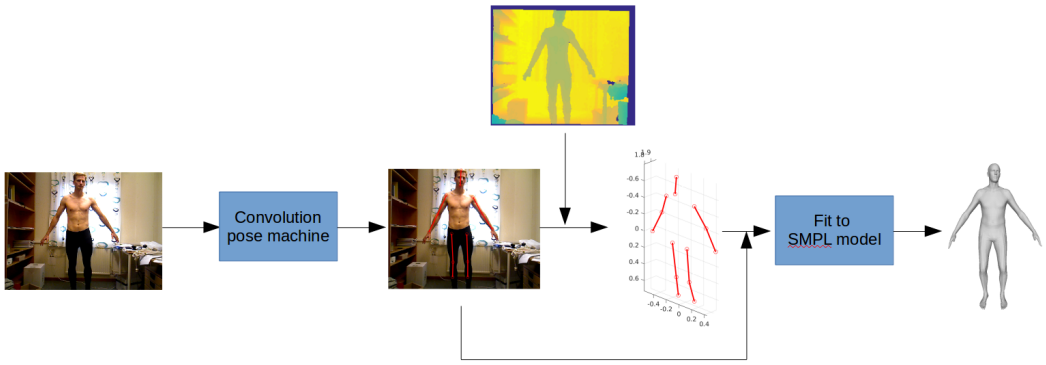


Figure 2: The overview of our method.

3 METHOD

The overview of our method is shown in Figure 2. We firstly use an RGB-D camera with known intrinsic parameters to capture an RGB image and a depth image. The skeleton is estimated by CPM, which designs a sequential architecture composed of convolutional networks to predict the 2D human pose. The corresponding 3D pose can be computed through the depth image and the intrinsic parameters of the camera using

$$X(i,j,z) = \left(\frac{(i-c_x)z}{f_x}, \frac{(j-c_y)z}{f_y}, z\right)^T, \quad (1)$$

where i, j are the coordinates of 2D joint points in the image plane, z is the corresponding depth for the point (i, j) and f_x, f_y, c_x, c_y are focal lengths and the optical center of the camera. Then, using 2D and 3D pose simultaneously, we build an objective function to measure the difference between the SMPL model and the observed person. Finally, the pose and shape parameters of the SMPL model can be estimated through minimizing the objective function.

3.1 The SMPL Model

As stated in (Loper et al., 2015), the SMPL model is a state-of-the-art human model for representing the pose and shape of a human body model. The model is a triangulated mesh with $N = 6980$ vertices and is defined as a function of shape and pose parameters. The shape parameters are represented as $\boldsymbol{\beta}$ and they are the coefficients of a linear shape space which is learned from a dataset containing thousands of registered scans. The pose parameters $\boldsymbol{\theta}$ are defined as the joint angle rotations. There are 24 joint points which contain 1 parent point and the pose of each joints is represented as vector in $\mathbb{R}^3$ encoding the relative rotation with respect to its parent in the kinematic tree. Therefore, the pose parameters have $23 \times 3 + 3 = 72$ elements. One of the advantages of SMPL model is that the position of skeleton joints is also defined as a function of the shape parameters, $J(\boldsymbol{\beta})$. Then, the global coordinates of the joint points which are represented as $R_{\boldsymbol{\theta}}(J(\boldsymbol{\beta}))$ can be calculated by considering the pose parameters $\boldsymbol{\theta}$. Using a perspective camera model, the joint points of the SMPL model in global coordinates can be projected to the image plane, which converts 3D joint points to 2D joint points in the image. Here we have to note that the order of the joint points in SMPL model is different with the 2D human pose estimated by CPM.

3.2 Objective Function

In this part we will give the definition of the objective function. In (Bogo et al., 2014), the objective function only takes 2D pose as prior knowledge. Although some methods (Xu et al., 2017) also consider 3D pose as the prior information, the data term in (Xu et al., 2017) is defined as an L_2-distance function, which is not stable and robust enough. The objective function in our method will combine both 2D and 3D pose to define the errors between joint points of the SMPL model and the joint points estimated by the CPM. The new data term can better represent the position of the joint points, especially when the human body has occlusions. Our experiments will demonstrate the performance of the method. The error function is defined by the Geman-McClure penalty function ρ (Geman and McClure, 1987) so that it can deal with the noisy estimates (Bogo et al., 2014; Bogo et al., 2015; Huang et al., 2017; Weiss et al., 2011). The Geman-McClure function is defined as

$$\rho_\sigma = \frac{e^2}{\sigma^2 + e^2}, \quad (2)$$

where e is the error which is defined by 2D and 3D poses and σ is a constant (set to 100 in our experi-

ments). In addition, regularization terms are also defined according to the prior information provided by the SMPL model to make the results more stable and natural. This gives the complete objective function in our method as

$$E(\boldsymbol{\beta},\boldsymbol{\theta}) = E_J(\boldsymbol{\beta},\boldsymbol{\lambda}) + \lambda_\alpha E_\alpha(\theta) + \lambda_\theta E_\theta(\boldsymbol{\theta}) + \lambda_\beta E_\beta(\boldsymbol{\beta}), \quad (3)$$

where $E_J(\boldsymbol{\beta},\boldsymbol{\lambda})$ is the data term combining 2D and 3D pose, $E_\alpha(\theta)$ and $E_\theta(\boldsymbol{\theta})$ are two pose penalty functions and $E_\beta(\boldsymbol{\beta})$ is the shape prior function. λ_α, λ_θ and λ_β are the weights for the pose and shape prior function.

The data term in our method is built based on (Bogo et al., 2014). The difference of our method is that we consider the 3D information in the data term. The function is defined as

$$E_J(\beta,\lambda) = \sum_{joints} \rho(\Pi(R_\theta(J(\beta))) - J_{2d}) + \lambda_{3d} \sum_{joints} \rho(R_\theta(J(\beta)) - J_{3d}), \quad (4)$$

where $\Pi(R_\theta(J(\beta)))$ denotes the projection of 3D joint points of the SMPL model by the camera Π. λ_{3d} is the weight of the data term using 3D pose and it will help to improve the robustness of the whole method. The distance between the two joint points is measured by a Geman-McClure penalty function. This new data term, taking into account 3D pose, provides more spatial information of joint points, which is useful when some body parts are occluded by other objects or the torso.

The regularization term is there to keep the pose and shape of the human body natural and probable. As shown in (Loper et al., 2015), SMPL is a statistical human body model which is learned from large datasets and it can provide strong prior information. We use some of the pose and shape penalty functions in (Bogo et al., 2014) in our method. For the pose regularization, the first part is to keep elbow and knees bending naturally and is defined as

$$E_\alpha(\theta) = \sum_i exp(\theta_i), \quad (5)$$

where $\boldsymbol{\theta}_i$ is the pose of the i-th joint. The second part for the pose penalty is defined as

$$E_\theta(\theta) = \min_j \left(-\log(c g_i \boldsymbol{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\theta,j}, \Sigma_{\theta,j})\right), \quad (6)$$

where $\boldsymbol{N}$ is a mixture of Gaussians function, g_i is the weight and c is a positive constant. The Gaussian function is fit to approximately 1 million poses which are obtained by fitting SMPL to the CMU marker data using MoSh (Loper et al., 2014). There are more details in (Bogo et al., 2014). For the shape regularization part, because the shape parameters of the SMPL model are obtained by training a dataset, the shape penalty function is also defined based on this

$$E_\beta(\beta) = \beta^T \Sigma_\beta^{-1} \beta, \quad (7)$$

where Σ_β^{-1} is a diagonal matrix and the elements on the diagonal are the squared singular values. These values can be estimated by Principal Component Analysis from the SMPL training shape set.

3.3 Optimization

The optimization has two steps. The first step is to estimate the camera translation. Here the focal length of the camera is known. The camera translation can be estimated through similar triangles defined by the torso length of SMPL and the predicted 2D pose.

The second stage is to fit the model through minimizing Eq. (3). The parameters for λ_θ and λ_β will decrease gradually during the optimization. Because the human body may either be facing the camera or not, we can firstly try an initialization and then rotate by 180 degrees. And we choose the one with smaller errors.

The procedure of minimization is implemented through Powells dogleg method which is provided by the python module OpenDR (Loper and Black, 2014) and Chumpy. For a single image with size 240×320, it takes about 3 minutes for the minimization on a desktop machine.

4 EXPERIMENTS

Our method is qualitatively and quantitatively tested on the synthetic dataset SURREAL (Varol et al., 2017), the real dataset Human3.6M (Ionescu et al., 2014) and some data acquired by Kinect. We compare our method with SMPLify (Bogo et al., 2014) which only used 2D joints to construct the data term. The setting of the parameters in our experiments were $\lambda_{3d} = 100$, $\lambda_\alpha = 10$. The optimization has four stages and in each stage the maximum iteration is 20. The weights λ_θ and λ_β for each stage are $(404, 404, 57.4, 4.78)$ and $(100, 50, 5, 1)$. For the Geman-McClure penalty function, the parameter $\sigma = 100$.

4.1 Results on SURREAL

SURREAL is a synthetic dataset containing more than 40,000 videos in which the human body models are generated based on the SMPL model. Each video consists of 100 frames and the person in the

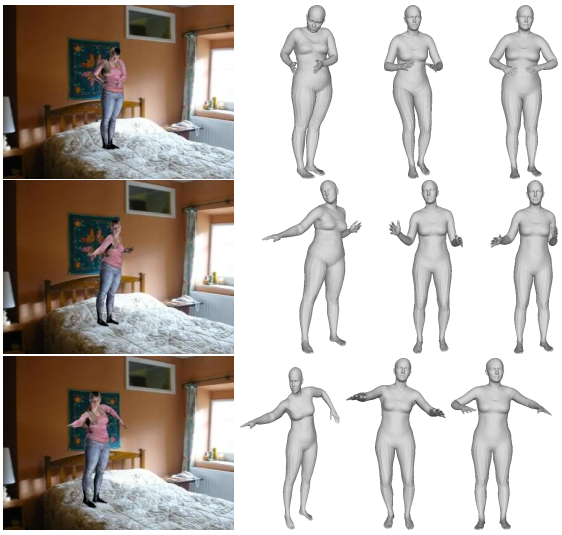


Figure 3: The results of some samples from the video in SURREAL. The first column is the sample from the video. The second column is the ground truth. The third column is the result of SMPLify. The last column is the result of our method.

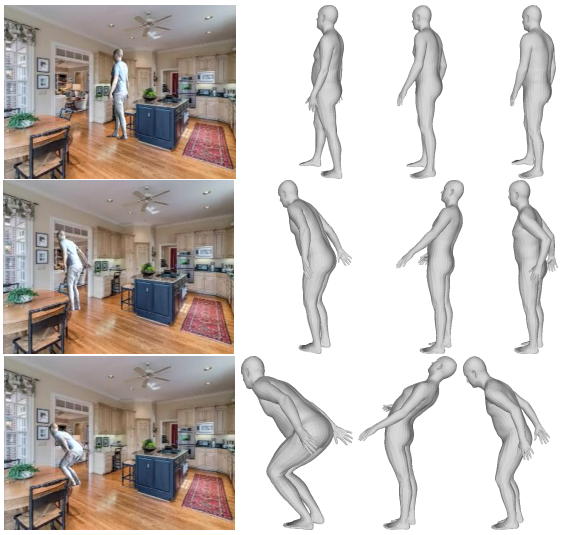


Figure 5: The results of some samples from the video in SURREAL. The first column is the sample from the video. The second column is the ground truth. The third column is the result of SMPLify. The last column is the result of our method.

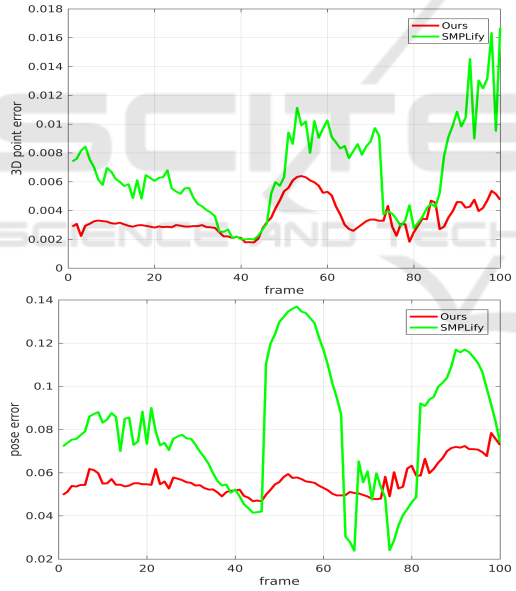


Figure 4: The error of the video in Figure 3. The left is the error of 3D vertexes between the results from our method and SMPLify and the ground truth.The right is the error of pose parameters between the results from our method and SMPLify and the ground truth.

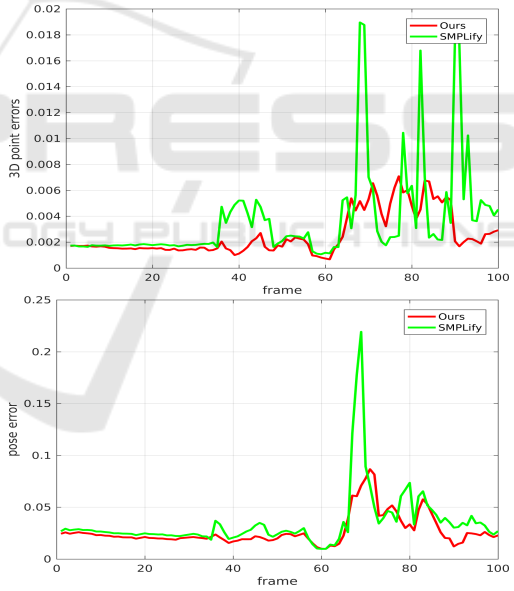
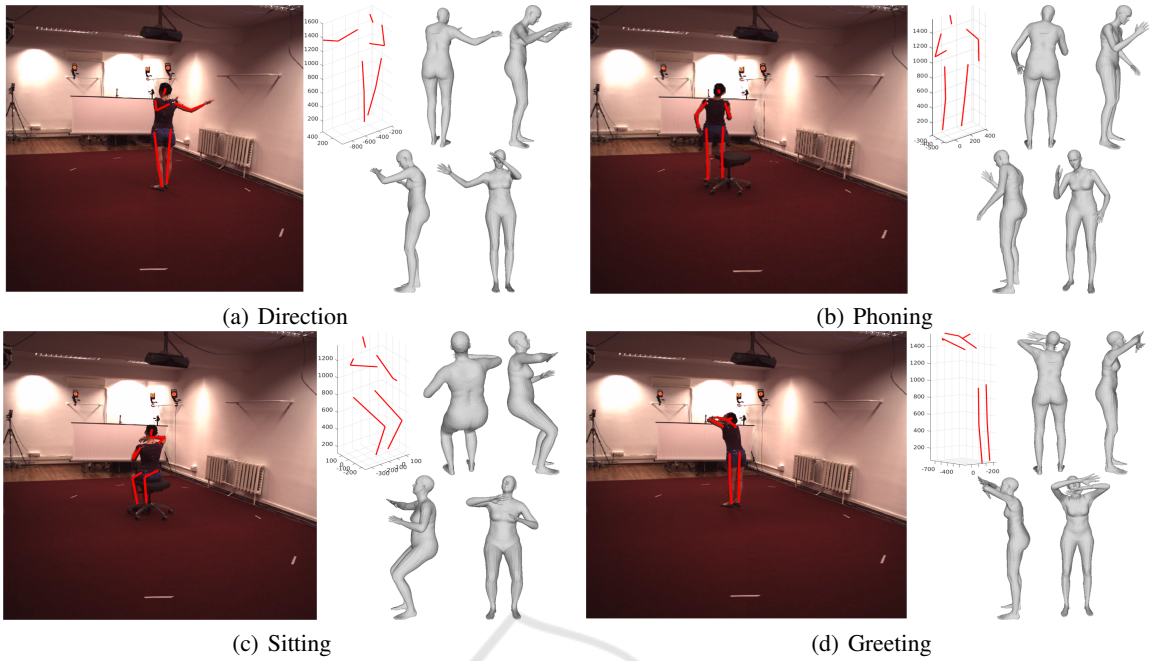


Figure 6: The error of the video in Figure 5. The top is the error of 3D vertices between the results from our method and SMPLify and the ground truth.The bottom is the error of pose parameters between the results from our method and SMPLify and the ground truth.

video has different actions and backgrounds. Especially, the ground truth of the pose and shape parameters of the person in this dataset are known and the 3D mesh of the person can be generated by using these parameters. Because the number of videos in the dataset is too large, we only use two of them to do the experiments. Figure 3 and 5 are qualitative results of the experiments. In the first column of Figure 3 and 5, three frames of the two examples are shown. The corresponding ground truth of the 3D mesh are given in the second column. The results of SMPLify and our method are shown in the third and fourth columns of Figure 3 and 5. The errors of the pose parameters and 3D vertices between



(a) Direction (b) Phoning

(c) Sitting (d) Greeting

Figure 7: The 3D model by our method from Human3.6M. For each subfigure the 2D and 3D joint points and 3D model from different views are given. There are 4 actions: (a) Direction; (b) Phoning; (c) Sitting; and (d) Greeting.

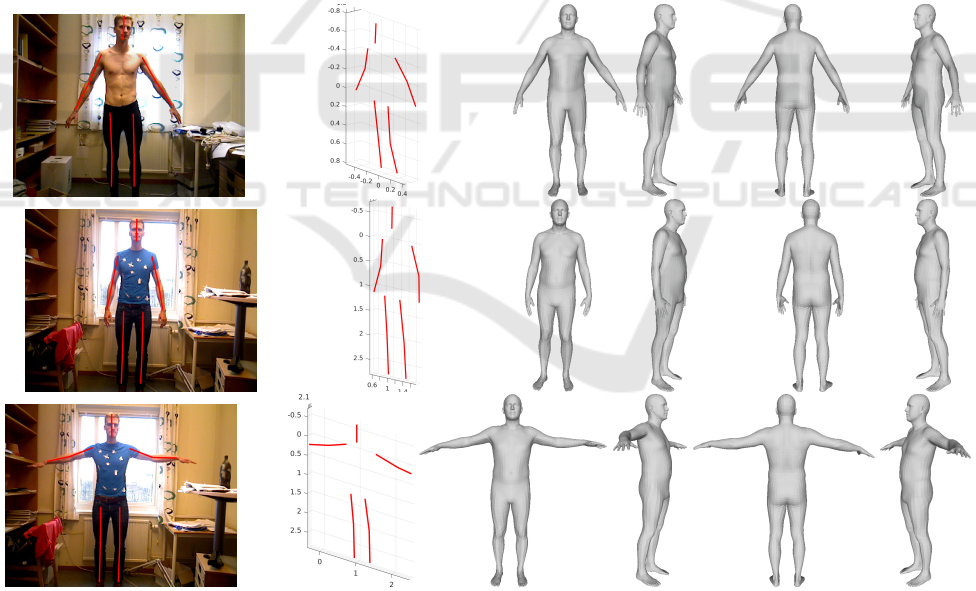


Figure 8: The 3D model of images acquired by ourselves estimated by our method. From left to right: the image with 2D joint points estimated by CPM, 3D joint points and 3D model from different view.

the ground truth and the estimated results are shown in Figure 4 and 6, which is the quantitative results. We can see from the two figures that the errors of our method are smaller than SMPLify in general. In addition, our method is more robust because the errors of each frame in the two videos are more stable when the 3D pose is used. Moreover, as shown in Figure 5, the results of SMPLify are obviously incorrect because the human body bends to another direction. This demonstrates that using only 2D pose can not provide enough space information for the algorithm. By contrast, our method can avoid this problem and obtain more accurate results because the 3D joints can provide more space information.

4.2 Results of Human3.6M

Human3.6M is a dataset containing 3.6 million 3D human poses and corresponding images. The whole dataset is obtained from four calibrated cameras which ensures the accuracy of the 3D joint positions. There are 11 subjects (6 males, 5 females) in the dataset and each of them perform 15 different actions. In this experiment we only give the qualitative results using some images from the first subject (Direction, Phoning, Sitting and Greeting) to test our method, since the dataset does not have the ground truth of pose parameters and 3D mesh. Figure 7 shows the reconstructed 3D models of subject 1 with different actions by our method.

We can see from Figure 7 that the 3D models estimated by our method for different actions roughly have the same pose as the person in the original images. These images are taken from the back of the person and some body parts are in the front of the torso. For example, the arms of the person in Figure 7 (a) and (c) are occluded by the torso, which makes it hard to to estimate the 3D model only using 2D information. Because the 3D joint points provide the spatial information, our method can roughly estimate the position of the arms. However, it also shows that the shape of the body is not perfect for the person in the original image. This is because the 2D and 3D joint points have stronger prior information for the pose but these cues do not provide enough shape information. We should also note that the pose of the hands in the 3D body model is not estimated well because our method does not consider the hand pose.

4.3 Results from our Own Data

Here we show results on RGB-D images acquired by Kinect. The person stands in front of the camera and has different poses. The intrinsic parameters of the camera used are $f_x = 262.5, f_y = 262.5, c_x = 159.75$, and $c_y = 119.75$. The 2D points are detected through the CPM model and the corresponding 3D joints are computed by considering the depth image using the intrinsic parameters.

Fitting results for the images are shown in Figure 8. The 2D joint points predicted by CPM are roughly correct from the first column in Figure 8. The 3D pose is computed based on the 2D pose and depth image and are also shown in the second column in Figure 8. The 3D model observed from several different views is also shown from the third to sixth columns in Figure 8. The 3D shape models are recovered from one single RGB-D image. Although only one image is used, the 3D model obtained by our method looks quite close to the action of the person in the original image.

5 CONCLUSIONS

We have presented a template-based 3D human body reconstruction method using a single RGB-D images. Our method considers both 2D joint points which can be estimated by a CNN-based model and 3D joint points which can be computed by the depth image. Then, these joint points are used simultaneously to fit a parametric human body model, SMPL. The pose and shape parameters can be estimated through minimizing the distance of joint points between observed images and the SMPL model. We test our method on the data taken by ourselves and the Human3.6M dataset. The results demonstrate that our method can estimate the 3D mesh of the observed images well.

In future work, more shape information should be considered to add to the objective function so that the final 3D model has good shape appearance. In addition, it is also possible to improve the accuracy of the reconstruction by using more images from different views.

ACKNOWLEDGEMENTS

This work was partially supported by the strategic research projects ELLIIT, eSSENCE, the Sten K. Johnson foundation and the China Scholarship Council (CSC).

REFERENCES

Alldieck, T., Magnor, M., Xu, W., Theobalt, C., and Pons-Moll, G. (2018). Video based reconstruction of 3d people models. In *arXiv preprint arXiv:1803.04758*.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416.

Ari, S., Andrew, F., Ruizhe, W., Hao, L., Mark, B., Gerard, M., and Evan, S. (2014). Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–211.

Bogo, F., Black, M. J., Loper, M., and Romero, J. (2015). Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *ICCV*, pages 2300–2308.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P. V., Romero, J., and Black, M. J. (2014). Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *ECCV*.

Cui, Y., Chang, W., Nöll, T., and Stricker, D. (2012). Kinectavatar: Fully automatic body capture using a single kinect. In *Computer Vision - ACCV 2012 Workshops*, pages 133–147, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S. R., Kowdle, A., Escolano, S. O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., and Izadi, S. (2016). Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4):114:1–114:13.

Geman, S. and McClure, D. (1987). Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 52:5–21.

Guo, K., Xu, F., Wang, Y., Liu, Y., and Dai, Q. (2015). Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *ICCV*, pages 3083–3091.

Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P. V., Romero, J., Akhter, I., and Black, M. J. (2017). Towards accurate marker-less human shape and pose estimation over time. In *3DV*.

Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., and Stamminger, M. (2016). Volumedeform: Real-time volumetric non-rigid reconstruction. In *ECCV*.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. (2011). Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 559–568.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018). End-to-end recovery of human shape and pose. In *CVPR*.

Leroy, V., Franco, J.-S., and Boyer, E. (2017). Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In *ICCV*.

Li, H., Adams, B., Guibas, L. J., and Pauly, M. (2009). Robust single-view geometry and motion reconstruction. *ACM Trans. Graph.*, pages 175:1–175:10.

Loper, M., Mahmood, N., and Black, M. J. (2014). Mosh: Motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6):220:1–220:13.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16.

Loper, M. M. and Black, M. J. (2014). OpenDR: An approximate differentiable renderer. In *ECCV*.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., and Theobalt, C. (2017). Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4).

Newcombe, R. A., Fox, D., and Seitz, S. M. (2015). Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*.

Pons-Moll, G., Romero, J., Mahmood, N., and Black, M. J. (2015). Dyna: A model of dynamic human shape in motion. *ACM Trans. Graph.*, 34(4):120:1–120:14.

Slavcheva, M., Baust, M., Cremers, D., and Ilic, S. (2017). KillingFusion: Non-rigid 3D Reconstruction without Correspondences. In *CVPR*.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. (2017). Learning from Synthetic Humans. In *CVPR*.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

Weiss, A., Hirshberg, D., and Black, M. J. (2011). Home 3D body scans from noisy image and range data. In *ICCV*.

Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H., and Theobalt, C. (2017). Monoperfcap: Human performance capture from monocular video. *CoRR*, abs/1708.02136.

Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., and Liu, Y. (2017). Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *ICCV*.

Zhang, Q., Fu, B., Ye, M., and Yang, R. (2014). Quality dynamic human body modeling using a single low-cost depth camera. In *CVPR*, pages 676–683.

Zollhofer, M., Niessner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., and Stamminger, M. (2014). Real-time non-rigid reconstruction using an rgb-d camera. *ACM Trans. Graph.*, 33:156:1–156:12.