3D Human Reconstruction using single 2D Image

Polina Katkova
Samara National Research University
Samara, Russia
Lin997@yandex.ru

Pavel Yakimov
Samara National Research University
Samara, Russia
yakimov@ssau.ru

***Abstract*—Computer Vision technology is rapidly developing nowadays. The need for 3D-reconstruction methods increases along with a number of Computer Vision system implementation. The highest need is for methods, which are using single image as an input data. This article provides an overview of existing methods for 3D-reconstruction and an explanation of planned implementation, which consists of a platform and a 3D-reconstruction algorithm using single image. Also, this article contains implementation of the Telegram bot, which allows anyone to test PIFu and an overview of the Mask R-CNN which will be used in this work later on.**

Keywords*—*3D Reconstruction, 3D Human body recovery algorithms, PIFu algorithm, Telegram bot, Segmentation methods, Mask RCNN

I. Introduction

At the moment using a 3D model instead of a real physical object often is a very important requirement. Digital copies provide more variety and flexibility to users than physical objects. Besides that, using a digital model can save a lot of time because it can be used despite the location of the model prototype and because the process of parameter calculating can be some degrees faster than while using a real model[1].

Computer Vision (CV) systems are widely used nowadays. Most of them have only one camera, so they are not able to capture a set of images from different angles. So, the possibility to create 3D content via a single image is getting highly relevant. The progress of such methods as deep learning, neural networks and segmentation algorithms helps to simplify the process of 3D reconstruction and thus, will help to develop different areas, such as CV or immersive technologies.

The range of 3D reconstruction method usage also contains the following areas: medicine (e.g. in computer tomography), Computer Vision (e.g. scene reconstruction, which can be used for calculating a trajectory of movement), microscopy, cinematography, multiplication, video-tracking (e.g. for biometric person identification), retail (e.g. online product demonstration in 3D), immersive technologies et cetera.

The article contains an overview of frameworks for popular 3D Human reconstruction methods. Most of those methods have been released in the past few years. The three following types of methods were considered: parametric methods, methods of recovering human shape and pose and human body recovery methods. Said methods use a combination of such methods as Convolutional Neural Networks, Semantic Segmentation, Marching cubes et cetera.

In the case, the input data consists of multiple images which have a different angle of view (an example of the process of getting an image set with different points of view is illustrated in Figure 1) the result of 3D reconstruction is almost unambiguous. Some years ago the company Autodesk released a new product named Recap, which is able to reconstruct a 3D model via image set [2]. However, in reality there is a higher need for 3D reconstruction methods via single image, because it has a higher practical use.

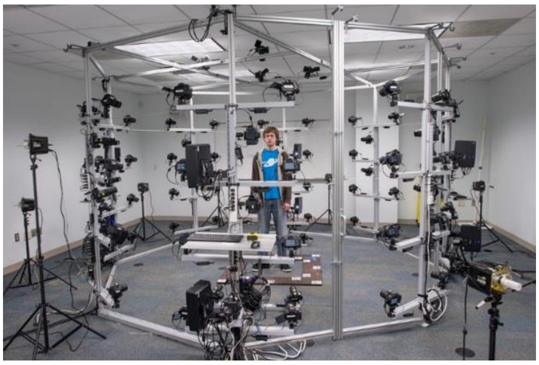


Fig. 1. Example of Studio, which allows getting a set of images from different angles [4].

The problem of model reconstruction via single image is an ambiguity of the back side shape definition (which is not visible on the picture) [3]. There is a similar problem with texturing – the texture part which is visible can be partially copied, but the reverse side has to be calculated by an algorithm which has to be implemented.

II. The overview of existing methods

A. *Algorithms for recovering human shape and pose*

Algorithm End-to-end Recovery of Human Shape and Pose was released in June 2019. It allows human model recovering via a single image. Unlike other methods, End-to-end Recovery can determine the location of key joints even if the person in the photo is turned away.

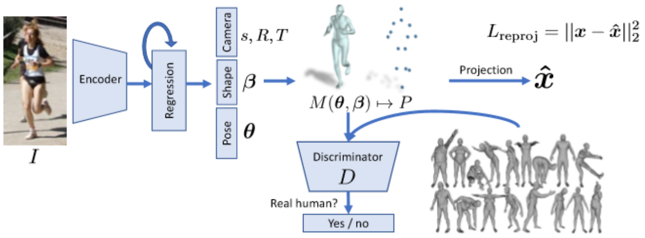
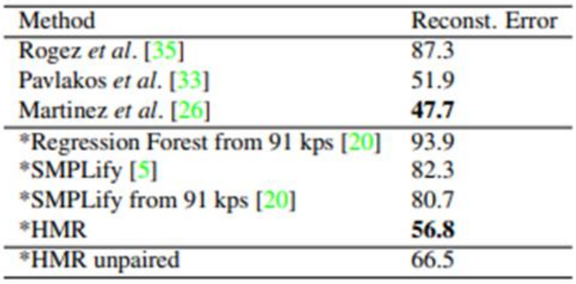


Fig. 2. Scheme of End-to-end Recovery of Human Shape and Pose algorithm [5].

The input data is an RGB image. Firstly, the image passes through a convolutional encoder and then the result is sent to the 3D regression module which iteratively minimizes the loss on the 3D model. Lastly the result passes the discriminant module, which determines if the resulting 3D model belongs to a person or not. The scheme of End-to-end Recovery of Human Shape and Pose algorithm is shown in the figure 2.

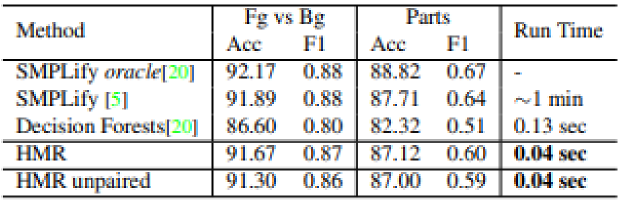
There was conducted a number of experimental studies about this method. The compaction of 3D reconstruction losses for different methods is illustrated in figure 3.



Method	Reconst. Error
Rogez *et al.* [35]	87.3
Pavlakos *et al.* [33]	51.9
Martinez *et al.* [26]	**47.7**
*Regression Forest from 91 kps [20]	93.9
*SMPLify [5]	82.3
*SMPLify from 91 kps [20]	80.7
*HMR	**56.8**
*HMR unpaired	66.5

Fig. 3. Comparation of HMR with other methods by criteria of 3D reconstruction loss [5].

The comparation of HMR with other methods by time executing is illustrated below, in the figure 4.



Method	Fg vs Bg		Parts		Run Time
	Acc	F1	Acc	F1	
SMPLify *oracle*[20]	92.17	0.88	88.82	0.67	-
SMPLify [5]	91.89	0.88	87.71	0.64	~1 min
Decision Forests[20]	86.60	0.80	82.32	0.51	0.13 sec
HMR	91.67	0.87	87.12	0.60	**0.04 sec**
HMR unpaired	91.30	0.86	87.00	0.59	**0.04 sec**

Fig. 4. Comparation of HMR with other methods by criteria of time needed for 3D reconstruction [5].

Illustrations in figures 3 and 4 show that the End-to-end Recovery has the best results in comparation with other methods.

A. *Algorithms for recovering human shape and pose*

The previous algorithm was able to recover the human shape, but not the shape of the clothes. The method *SiCloPe: Silhouette-Based Clothed People* was released in august 2019 and it has the ability to reconstruct human shapes and clothes. After the 3D model reconstruction process, *SiCloPe* recreates the model texture.

The algorithm consists of the following steps: firstly, it defines 2D human silhouettes and creates a 3D map with model joint locations; secondly, the method generates new 2D silhouettes of the model via a 3D joint location map; after that, SiCloPe reconstructs the 3D model by using a set of 2D silhouettes from step 2. If the 2D silhouettes are built incorrect then the grid used for reconstruction also will not match the actual model. SiCloPe uses an algorithm of deep surface recognition, which includes “greedy sampling”. Using this algorithm guarantees that the reconstruction grid will be correct. The last step of the algorithm is texturing the reconstructed model.

The scheme of the SiCloPe algorithm is illustrated below, in the figure 5.

Method PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization [7] was released in November 2019. This method allows reconstructing 3D models by using one image or a set of images. The feature of PIFu is a high-quality texture reconstruction even on the invisible parts of the object in the picture.

The algorithm is able to reconstruct even complicated figures, which includes crumbled clothes, high heels or complex hair-style.

The algorithm PIFu consists of a convolutional encoder and a continuous function. The overview of PIFu’s framework is illustrated in the figure 6.

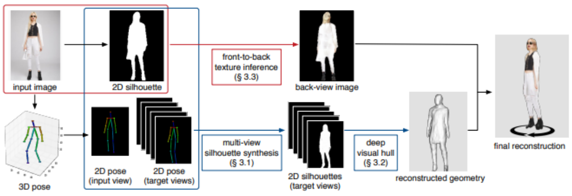


Fig. 5. Overview of SiCloPe: Silhouette-Based Clothed People’ framework [6].

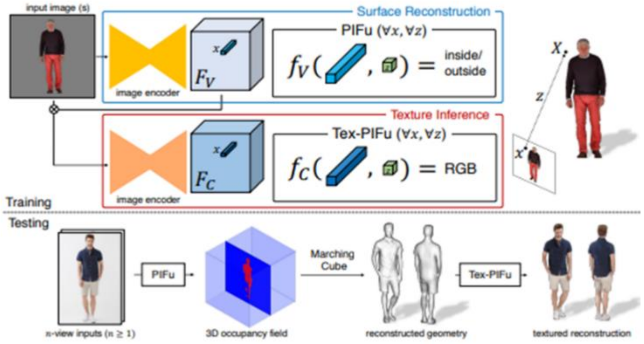


Fig. 6. The overview of PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization [7].

B. *The parameterized algorithms of human recovery*

An algorithm named skinned multi-person linear model (SMPL) [8] is one of the most popular parameterized algorithms of human body recovering. SMPL was released in 2015 and it is still being used in other 3D reconstruction works as part of the implementation or for a comparation process.

SMPL has been trained on some thousands of 3D models of human bodies which have different forms and figures. The recovered 3D model has a map with data of weight at each point of the body model, so the joints can look realistic when a model is changing its pose.

3D models recovered via SMPL algorithm can be used in such programs as Autodesk Maya or Unity, where they can get animated later on.

The SMPL model is illustrated below, in the Figure 7.

Another popular parametric algorithm is Shape Completion and Animation of People (SCAPE) [9]. SCAPE was published in 2005 in the ACM Transactions on Graphics journal.

SCAPE allows to combine a single scan of a person with a motion markers sequence. So, as the result this algorithm returns an animation made by mixing a body shape with a pose.

The algorithm consists of three parts: pose deformation, body shape deformation and animation via motion capture data. The body shape can get deformed by changing a template shape with four possible parameters, such as height, weight, muscularity and gender. The overview of the deformation parameters are illustrated in the Figure 8. In the case the body scan is missing a part of a surface, the SCAPE can complete the shape using the Correlated Correspondence (CC) algorithm [10]. The pose can be deformed via CC algorithm as well.

The authors have created two data sets: the pose data set consists of 70 poses and the shape data set, consists of 45 different body shapes. Also, the SCAPE algorithm can be applied to other shapes than human.

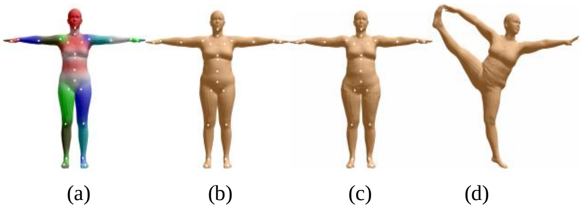


Fig. 7. SMPL model: (a) – human model with a weight grid, (b) – parametrized human model, (c) – human body model with mixed shape on, (d) – human body model in a pose [8].

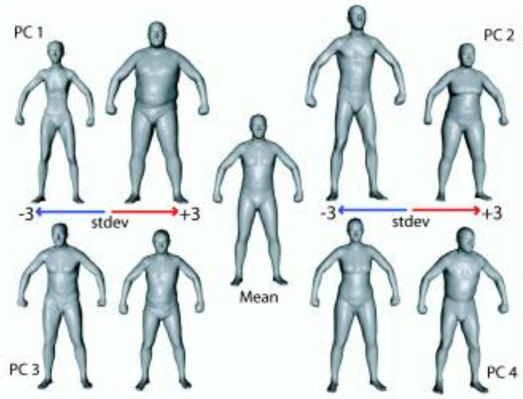


Fig. 8. The four parameters for shape deformation in the SCAPE model.

C. *The overview results*

The future purpose of current work is creating a virtual fitting room. It is proposed to use the PIFu algorithm for this purpose. The reason for this choice is an open repository and a simple installation and run of PIFu. The Implementation of PIFu uses an RGB image of human body and a mask which allows to detect a human on the image. It is proposed to research segmentation method Mask R-CNN and implement it for future realization, so the only image can be used as an input data for the PIFu algorithm. This method will be overviewed in the next chapter.

III. Experimental Research

The PIFu algorithm has been tested on different data while experimental studies. The input data consists of a photo and a created mask for this image (to determine an object on the background) via Photoshop. The images have PNG format and a resolution of 720x1080 pixels. The input data and the results for each of the three experiments are illustrated in Figures 9-12. In the Figure 12 is demonstrated that the result of the fourth experiment is not very precise and has a high loss.

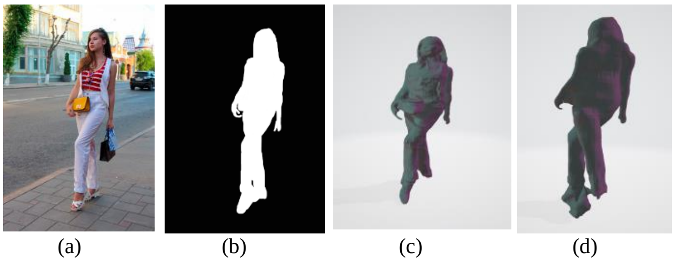


Fig. 9. Results of the first experiment: (a) – input RGB image, (b) – mask for the Input image, (c) – front view of the resulting 3D object, (d) – view of the resulting 3D object from the backside.

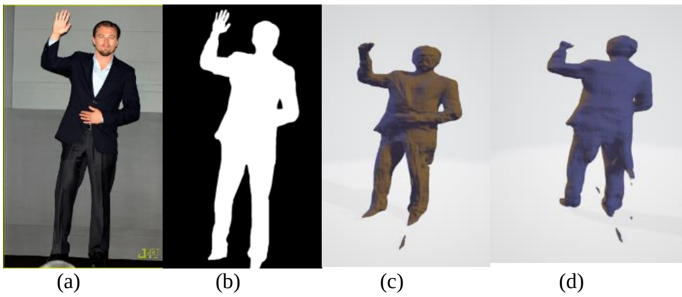


Fig. 10. Results of the second experiment: (a) – input RGB image, (b) – mask for the Input image, (c) – front view of the resulting 3D object, (d) – view of the resulting 3D object from the backside[11].

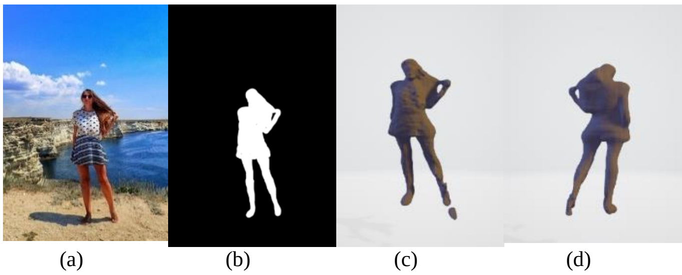


Fig. 11. Results of the third experiment: (a) – input RGB image, (b) – mask for the Input image, (c) – front view of the resulting 3D object, (d) – view of the resulting 3D object from the backside.

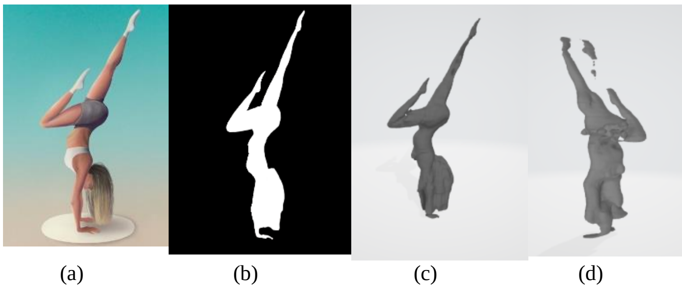


Fig. 12. Results of the fourth experiment: (a) – input RGB image, (b) – mask for the Input image, (c) – front view of the resulting 3D object, (d) – view of the resulting 3D object from the backside.

The run time of PIFu algorithm in the first experiment equals 8.92 seconds, in the second – 10.47 seconds, in the third – 7.19 seconds, and in the fourth – 7.34 seconds.

There are more result images on the following GitHub account: https://github.com/thePolly/PIFu. This repository contains the code for Telegram bot and PIFu's algorithm as well.

IV. Implementation

A. *Proposed implementation*

It is proposed that in this work a 3D human model recovery algorithm via single image has to be implemented. This algorithm can be used for virtual fitting room implementation later on. As an example for implementation, the earlier on overviewed methods can be used.

This method will consist of two convolution encoders for both 3D model and texture reconstruction.

It is planned that for the realization a dataset is created, which includes a set of 2D people images and a set of 3D models for these images. A Microsoft Kinect 2.0 camera and stereo-cam ZED 2K will be used for creating 3D objects. The Microsoft Kinect camera uses an infrared laser for determining the depth of the image matrix. The optimal distance between objects and the Kinect camera is between one and four meters [12]. Unlike Kinect, the ZED camera has no infrared sensor. ZED uses methods, which include artificial intelligence for determining the image depth [13].

A. Telegram bot

Anyone can use the Telegram bot as a platform for 3D reconstruction. Currently, the bot accepts a single image and a mask of this image as an input data and returns a file with resulting 3D model. For more details "/help" command can be used. The name of the bot is @human_body_recnstruction_bot.

B. Image segmentation

To make the process of using Computer Vision easier, a segmentation method has to be implemented. It will allow users to upload only one RGB image without any mask.

Mask R-CNN segmentation method has been released in 2018 by Facebook AI Research [14]. The framework allows to detect multiple number of objects of different type on the image.

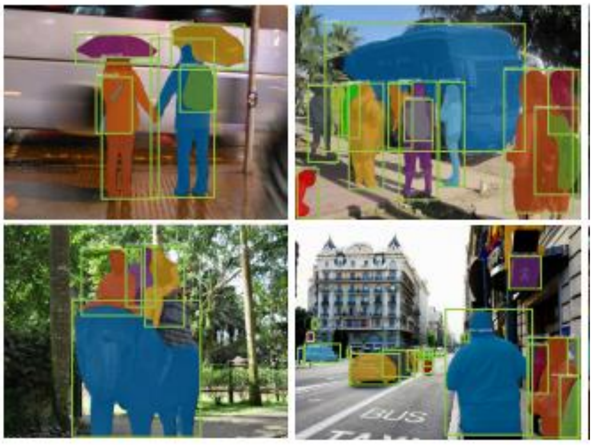


Fig. 13. Mask R-CNN results on the COCO test set [14].

The Mask R-CNN framework is based on the Faster R-CNN. The Faster R-CNN has two outputs, a class label and a bounding-box offset and the Mask R-CNN has an additional third mask output, which predicts the layout of the segmentation mask for detected object. So, the loss for the Mask R-CNN is defined as sum of losses for each output:

$$L = L_{cls} + L_{box} + L_{mask}, \quad (1)$$

where L_{cls} is classification loss, L_{box} is a bounding-box loss and L_{mask} is a mask definition loss.

The framework allows to choose specific classes to detect. For online fitting room implementation, the class list should contain only class for human bodies. The examples of Mask R-CNN detection are illustrated in the figure 13.

V. Conclusion

Thus, the following 3D reconstruction method types have been overviewed: recovering of human shape and pose, human model recovering and parametrized human recovery. Most of those methods can accept a single image as an input data.

The overview contains a description of the most popular 3D Human reconstruction methods. Each overview describes methods which have been used in the implementation process. This paper may help in the design phase of the method developing. It can be used to understand which type of 3D reconstruction method has to be implemented depending on the task and which technologies this method should include. Thus, to implement a virtual fitting room, it is appropriate to use a parametric method or a method of recovering human shape and pose, because then the resulting object will contain no clothing items.

In conclusion, the Telegram bot, which allows to test PIFu algorithm has been created. The proposed realization has been set. So, to create a virtual fitting room, firstly a dataset has to be filed, a method for 3D human pose and shape has to be implemented and the overviewed Mask R-CNN has to be implemented as well.

References

[1] O.V. Evseev, "Implementation and research of models and algorithms of 3D reconstruction of cloud points defined by a sequence of parallel sections," Ph. D, 2016.

[2] "Autodesk ReCup," Autodesk Knowledge Network, 2019.

[3] W. Choi, "Understanding indoor scenes using 3D geometric phrases," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 33-40, 2013.

[4] Photogrammetry, 2019 [Online]. URL: https://imgur.com/gallery/yuEncdf/comment.

[5] A. Kanazawa, "End-to-end recovery of human shape and pose," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122-7131, 2018.

[6] SiCloPe: Silhouette-Based Clothed People, arXiv Preprint: 1901.00049v2.

[7] Sh. Saito, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," Proceedings of the IEEE International Conference on Computer Vision, 2019.

[8] M. Loper, "SMPL: A skinned multi-person linear model," ACM transactions on graphics (TOG), vol. 34, no. 6, pp. 248, 2015.

[9] D. Anguelov, "SCAPE: shape completion and animation of people," ACM SIGGRAPH, pp. 408-416, 2005.

[10] D. Anguelov, "The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces," Advances in neural information processing systems, 2005.

[11] Leonardo DiCaprio Club, 2020 [Online]. URL: https://ru.fanpop.com/ clubs/leonardo-dicaprio/images/10841990/title/leonardo-dicaprio-photo.

[12] How stuff works. How Microsoft Kinect Works, 2020 [Online]. URL: https://electronics.howstuffworks.com/microsoft-kinect1.htm.

[13] ZED 2. Stereolabs, 2020 [Online]. URL: https://www.stereolabs.com /zed-2.

[14] K. He, "Mask r-cnn," Proceedings of the IEEE international conference on computer vision, 2017.