Overview of Methods for 3D Reconstruction of Human Models with Applications in Fashion E-commerce

Ivaylo Vladimirov[1], Desislava Nikolova[2] and Zornitsa Terneva[3]

Abstract – **In this scientific paper, an overview of different methodologies and algorithms used for the reconstruction of 3D human models from 2D videos of people in action is presented. These methods could be applicable in the growing e-commerce business. Due to emerging challenges of global warming and the coronavirus pandemic, many developments in the apparel industry must be made in order to make the branch more sustainable and eco-friendly. In this time of globalisation, many brands produce and sell their items internationally and use long-distance shipping to distribute and deliver to their shops and clients. With online shopping, there is a big concern with clothing being the "wrong size" or "wrong item". To tackle this problem companies, offer their clients the opportunity to return ill-fitting items, but that is a problem in itself, because it increases their carbon footprint and it creates unnecessary pollution. A possible solution may come from the use of a computer vision application – the development of a system for size estimation or trying on clothes virtually.**

Keywords – **Overview, 3D Reconstruction, E-commerce, Monocular Video, Image Processing;**

I. Introduction

For almost a century, the apparel industry has been an integral part of the retail business and a major source of pollution. For the fashion trade, 2020 was the year in which everything transfigured as the coronavirus pandemic sent shock waves around the world. Both "The State of Fashion 2021" report [1] and "2020 Online Apparel Report" [2] elucidated that consumer behaviour shifted majorly from traditional shopping to online shopping, not only because of the imposed worldwide quarantine, but also as a trend in the last decade [3]. At the end of 2020 e-commerce made up 39% of all apparel sales which is a drastic increase from 2019 – 27%. Figure 1 represents the percentage change of online retail sales year-on-year in the fashion industry, where 0% means that there is no increase and 100% that the percentage has doubled. [4]

Although e-commerce may seem more favourable and convenient for the customer, it poses more risks for the seller since returns, due to the receipt of articles that are the wrong size or just different, may become a definite concern. Such a situation would lead to the completion of three parcel transitions (1st: from the seller's storage to the client's address; 2nd: from the client's home back to the seller; 3rd: again from the seller to the client) and twelve courier trips. The itinerary of the courier's trips can be seen in Figure 2 where you can track the movement of a parcel: 1 to 4: Delivery, 5 to 8: Return, 9 to 12: Delivery of the right order. In addition, a lot of paper and plastic is used for packaging and distribution and then thrown away, making added unnecessary debris.

Fig. 1 Percentage change of online retail sales year-on-year from 2005 until 2020.

Most fashion brands and shops currently try to find the appropriate sizing by asking a customer his/her/their preferences from favourite or often worn apparel. A superior solution for this problem is to give the ability to try clothes on virtually to a customer. Developing an optimized algorithm is a good way to reduce shipping's carbon footprint up to 2 times and create a new profitable experience. [5]

The main goal is to create an intelligent system that uses a 2D video to spawn a 3D model of a real person (client of a fashion brand). This model will be utilized to visualize selected clothes onto the person and estimate the best size for him/her/them. A personalized realistic and able to be animated 3D model of a human can also be used for many other applications, including virtual and augmented reality, human tracking for surveillance, gaming, or biometrics, entertainment, health-care and many more.

[1]Ivaylo Vladimirov is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: ivladimirov@tu-sofia.bg.

[2]Desislava Nikolova is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria. E-mail: dnikolova@tu-sofia.bg

[3]Zornitsa Terneva is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria. E-mail: zterneva@tu-sofia.bg

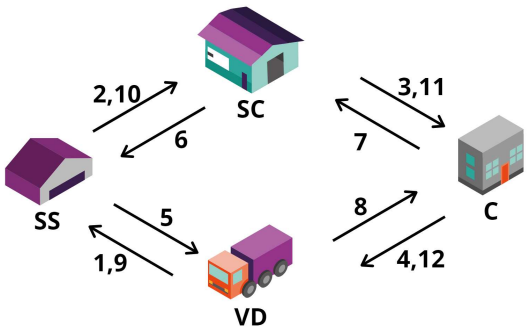

Fig.2 The route of the courier. SS - Seller's Storage, SC – Sorting Centre, C – Customer, VD – Vehicle Depot

II. METHODS CLASSIFIED BY NUMBER OF STAGES

The advent of powerful deep convolutional neural networks (DCNN) and the development of large-scale datasets [6] allowed significant progress in 3D human pose estimation from 2D videos. Recent advancements in deep learning have shifted focus on the creation of more challenging solutions to the problem: monocular 3D reconstruction. Monocular reconstruction inherently is an ill-posed problem which introduces a few more challenges such as considerable viewpoint fluctuation, self-occlusions and obstructions due to free-form attire. This area's approaches can roughly be divided into two categories:

A. One-stage approaches

In this type of approaches the 3D models are built directly from a traditional two-dimentional image. One of the first concepts, proposed by Li et al. [7], uses a DCNN that is trained simultaneously via a multi-task framework throught pose joint regression and detection of body parts. For modeling high-dimensional joint dependencies, Tekin et al. [8] proposed the use of an additional autoencoder. Pavlakos et al. [9] suggested a voxel relaxation technique that replaces joint coordinates with voxel representation, as well as a coarse-to-fine learning strategy, to reduce the regression computation time. Large-scale 2D pose datasets are not suitable for these methods since they need extensive annotation for the training.

B. Two-stage approaches

Two-stage approaches utilize up-to-date 2D pose evaluators, which removes dataset requirements, by first making an estimation on the monicular pose and lifting it to a 3D one and then predicting the locations on the 3D model by building the depth regression. [6]. This approach is represented in the work of Zhou et al. [10] where a weakly supervised learning technique has been incorporated. The technique is based on the function of geometric loss used for the depth regression module's training. Two more examples are the algorithms proposed by Martinez et al. [11] and Moreno-Noguer [12] where they respectively use a fully connected residual network and a pairwise distance matrix as the depth regressor.

III. METHODS CLASSIFIED BY METHODOLOGY

Model reconstruction of people in clothing can also be categorized according to two additional criterias: the camera/sensor type and the kind of template used for reconstruction. When working with inputs from depth and multi-view cameras or fusion of sensors it is typical to use free-form methods. These methodologies quite precisely reconstruct surface geometry without the need for a good prior awareness of the form. If the use of a template is involved the method can be characterized as model-based. These approaches, such as the SMPL algorithm, aim to restore the 3D surface model by fitting a parametric body representation. Model-based techniques accurately approximate the form and posture of the underlying naked body, but they fall short of reconstructing fine surface structure specifics of the body and wrapped clothes.

A. Free-form methods

Free-form techniques distort a mesh or use a volumetric approximation of shape to recreate the moving geometry. Though modular, such methods necessitate high-quality multi-view input data, making them unfit for many implementations. By iteratively fusing geometry in a canonical frame, systems like KinectFusion can recreate 3D rigid scenes and appearance models using a depth camera (Zhou et al.[13]). Cui et al. [14] and Shapiro et al. [15] propose algoritms that adapt KinectFusion for human body scanning. The issue is that these methods require different footage shot for distinct periods. As a consequence, the subject/client/person must either stand still while the camera rotates or subtlely change pose. Using multiple kinects or multi-view represented in the techniques of Dou et al.[16], Leroy et al. [17] and Orts-Escolano et al. [18] is one way to make fusion and monitoring more stable. While these approaches yield excellent reconstructions, they do not register all frames to the same template and are focused on different applications. The problem of tracking non-rigid deformations is minimized by pre-scanning the object or individual to be monitored. Xu et al. [19] used a pre-captured outline design to recreate human pose and distort textile geometry from monocular footage. They also enhance shape prediction by aligning visual hulls over time. In the articulated example, they must segment and monitor each body part separately before combining the data in a coarse voxel model; more specifically, multi-view feedback is required. This approach has the advantage of allowing the reconstruction of general dynamic forms if a reference surface is available.

B.Model-based methods

A parametric body model is used in many works to estimate human posture and form from photographs. Easy primitives were used to construct early computer vision models. Recent ones encode posture and shape deformations by learning from thousands of scans of actual humans. Some studies use temporal knowledge to recreate body shape from depth data sequences. To take advantage of the temporal detail, a single shape and multiple poses are usually optimized. Using multi-

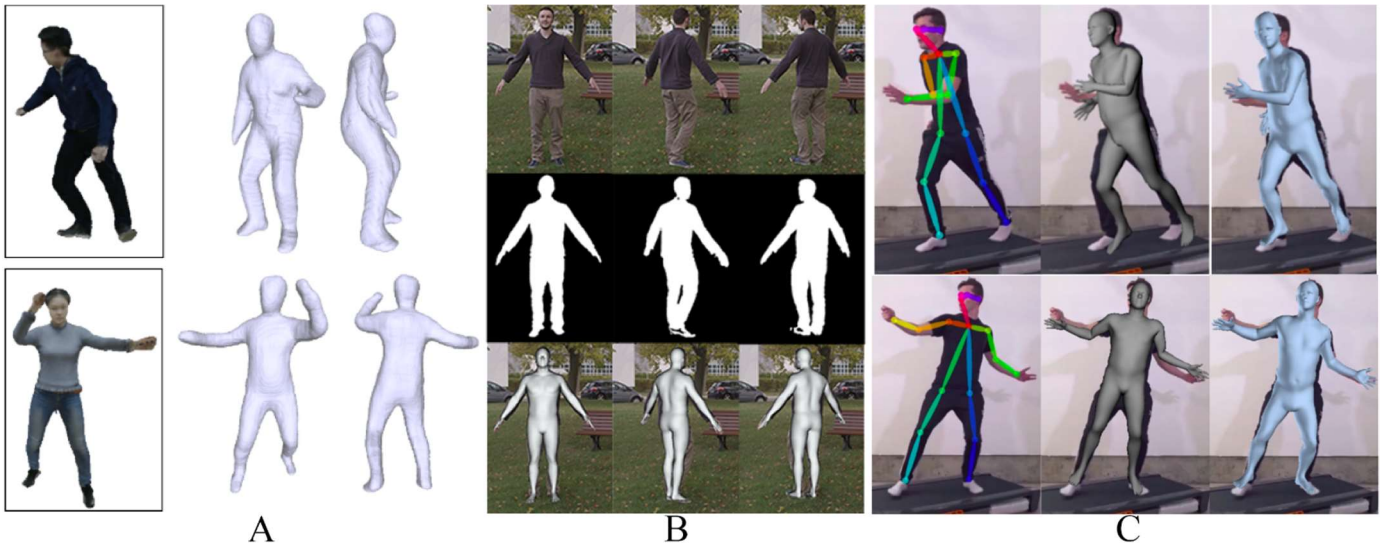

Fig.3 Results from the 3D Reconstruction of Models from monocular images and side-by-side comparison of pose estimation: A – from a single RGB image [21]; B – from a pre-recorded monocular video [24]; C - from a real-time stream of video data [25].

view, several researchers have demonstrated outdoor output capture [20] using a total of Gaussians body model or a pre-computed prototype. A variety of studies have focused on inferring the shape parameters of a body model [21] from a single picture using silhouettes, shading signals, and colour. In difficult cases, advances in 2D pose identification have made 3D pose and shape prediction feasible. Lassner et al. [22] match a 3D body model to 2D detections and since only model parameters are optimized, these methods depend heavily on 2D detections, the results are similar to the shape space mean.

IV. Key Findings

As an outcome of the overview, three algorithms stand out above the rest. They differ drastically in the input data they use. One uses a single RGB image [21], another a pre-recorded monocular video [24], and the last - a real-time stream of video data [25].

- A 3D human body model is encoded as a series of PeeledDepth & RGB maps by Jinka et al. [21]. They begin by considering the human body to be a non-convex entity in a virtual scene. A series of rays emanating from the camera center are traced through each pixel to the 3D world using a computer camera. The series of first ray intersections with the 3D body is saved as a depth map, which captures visible surface information nearest to the camera. The occlusion is then stripped away, and the rays are extended past the first bounce, reaching the next intersecting surface. Four intersections with each ray are taken into account - 4 Peeled Depth & RGB maps - in order to diligently reconstruct a human body, assuming that the system may manage self-occlusions induced by the most common body poses. Using traditional camera projection techniques, a point cloud can be generated from these charts. PeelGAN is a conditional GAN function generated by the authors to produce Peeled maps from an input image. PeelGAN produces Peeled Depth maps and matching RGB maps from a single RGB image as input. The initial RGB image is first fed to a feature map recovery encoder network, which consists of a few convolutional layers, before being fed to a series of ResNet [23] blocks. Since the Peeled Depth and RGB maps is sampled from various distributions, the machine uses Look-Up tables as its activation mechanism and decodes them in two separate divisions. Three Peeled RGB maps and four Peeled Depth maps are generated. (see Figure 3.A)
- The second method proposes a framework for making a personalized 3D human body prototype from a single moving human video. [24] The reconstruction includes customized hair, body, and wardrobe geometry, as well as surface texture and an underlying model that allows for adjustments in posture and form. It incorporates a refinement-oriented parametric human body model with surface displacements, as well as a novel approach for morphing and fusing complex human silhouette cones in a traditional frame of reference. The fused cones combine the shape details from the film, enabling us to refine the shape of a complex model. This algorithm not only captures the surface geometry and appearance, but it also rigs the body model with a kinematic skeleton, allowing for approximate pose-dependent surface deformation. Quantitative findings (see Figure 3.B) reveal that this method can recreate human body shape with a 4.5mm precision, and an ablation study demonstrates robustness to noisy 3D pose predictions.
- In the third algorithm [25] the authors developed a platform for real-time garment overlay. As a modification to Kanazawa et al. [26] algorithm for quick SMPL body mesh optimization, they produced a 3D uplifting model capable of competing with state-of-the-art methods. An iterative closest point technique using silhouette derived with the DeepLabV3 model was used to overcome potential objects in the 3D to 2D transformation. (see Figure 3.C) The ability of the method to perform all mapping computations on computers without dedicated GPUs, such as smartphones, was the key achievement.

For a visual comparison check figure 3, where output results for the three different approaches are presented. On the left is the input image and on the right the output 3D models.

V. Conclusion

As a conclusion a few possible optimization are given. The encoding in [21] is resistant to extreme self-occlusions while being reliable and effective in terms of learning and inference time. Few surface triangles that are tangential to the perspective of the input image are missed by the peeled representation. However, when building meshes from the corresponding projected point clouds, this constraint can be resolved with limited post-processing.

In [24] the framework can be further optimized for more speed or accuracy. Game engines like Unity7 or Unreal Engine 48 can be used to supplement 3D modelling software since they are optimized for real-time rendering. Alternatively, rather than processing each frame, several can be skipped and mapping can be created for them using animation interpolation. This will allow for the use of more sophisticated approaches for textile simulation, resulting in higher quality representation.

The last method [25] finds its limitations in its presentation of parts that do not share the same topology as the body: long hair or voluptuous clothing that can not be modelled as an offset from the body. Furthermore, it can only capture surface information visible on at least one view's outline. This means that concave areas like the armpits and inner thighs are often neglected. Quick skeletal motions can trigger strong fabric movement, which will result in a loss of detail. To allow a more realistic rendering and video augmentation, the method can integrate illumination and material estimation, as well as temporally changing textures.

Acknowledgement

This work is conducted under the grant of the National Program "Young Scientists and Postdoctoral Students", PMC №577, 2018-2021, Ministry of Education and Science, Bulgaria.

References

[1] I.Amed, A.Balchandani, A.Berg, S.Hedrich, J.E.Jensen, F.Rölkens, “The State of Fashion 2021”, BOF and McKinsey Company, 01.2021;

[2] “2020 Online Apparel Report”, Digital Commerce 360 Company, 06.2020;

[3] S.V.Kushwah, A.Singh, “From Traditional Shopping to Online Shopping A Study of the Paradigm Shift in Consumer”, Journal of General Management Research, Vol. 6, Issue 1, pp. 1–13 01.2019;

[4] W.Richter, “Online Sales by Category, in Weirdest Economy Ever”, WOLF STREET, 11.2020;

[5] S.Kent, I.Amed, “The sustainability gap”, Report of Business of Fashion, 03.2021;

[6] F.Bogo, A.Kanazawa, C.Lassner, P.Gehler, J.Romero and M.J.Black., ”Keep it smpl: Automatic estimation of 3d human pose and shape from a single image.,” , European Conference on Computer Vision, Springer, Cham, pp. 561-578, 2016;

[7] S.Li, A.B.Chan. “3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network”, Asian Conference of Computer Vision, 2014;

[8] B.Tekin, I.Katircioglu, M.Salzmann, V.Lepetit, P. Fua, “Structured prediction of 3d human pose with deep neural networks”, arXiv preprint, 2016;

[9] G.Pavlakos, X.Zhou, K.G.Derpanis and K.Daniilidis, ”Coarse-to-fine volumetric prediction for single-image 3d human pose”, IEEE Conference onComputer Vision and Pattern Recognition, 06.2017;

[10] X.Zhou, Q.Huang, X.Sun, X.Xue, Y. Wei, ”Towards 3d human pose estimation in the wild: a weakly-supervised approach”, IEEE Conference on Computer Vision, pp. 398-407, 2017;

[11] J.Martinez, R.Hossain, J.Romero, J.J.Little, ”A simple yet effective baseline for 3d human pose estimation”, IEEE Conference on Computer Vision, pp. 2640-2649, 2017;

[12] F.Moreno-Noguer, ”3d human pose estimation from a single image via distance matrix regression”, IEEE Conference on Computer Vision and Pattern Recognition, pp. 2823-2832, 2017;

[13] Q.Y.Zhou, V.Koltun, “Color map optimization for 3d reconstruction with consumer depth cameras”, ACM Transactions on Graphics, 2014;

[14] Y.Cui, W.Chang, T.N’oll, D.Stricker, “Kinectavatar: fully automatic body capture using a single kinect”, In Asian Conference on Computer Vision, Springer, 2012;

[15] A.Shapiro, A.Feng, R.Wang, H.Li, M.Bolas, G.Medioni, E.Suma, “Rapid avatar capture and simulation using commodity depth sensors”, Computer Animation and Virtual Worlds, 2014;

[16] M.Dou, S.Khamis, Y.Degtyarev, P.Davidson, S.R.Fanello, et al, "Fusion4d: Real-time performance capture of challenging scenes", ACM Transactions on Graphics,2016;

[17] V.Leroy, J.S.Franco, E.Boyer, "Multi-View DynamicShape Refinement Using Local Temporal Integration", IEEE Conference on Computer Vision, Italy, 2017;

[18] S.Orts-Escolano, C.Rhemann, S.Fanello, W.Chang, A.Kowdle, et al, "Holoportation: Virtual 3d teleportation in real-time", Symposium on User Interface Software and Technology,2016;

[19] W.Xu, A.Chatterjee, M.Zollhoefer, H.Rhodin, D.Mehta, H.P.Seidel, C.Theobalt, "Monoperfcap: Human performance capture from monocular video", ACM Transactions on Graphics, 2018;

[20] H.Rhodin, N.Robertini, D.Casas, C.Richardt, H.P.Seidel, C.Theobalt, “General automatic human shape and motion capture using volumetric contour cues”, European Conference on Computer Vision, pages 509–526. Springer, 2016;

[21] S.S.Jinka, C.Rohan, A.Sharma, P.Narayanan. “PeeledHuman: Robust Shape Representation for Textured 3D Human Body Reconstruction.” , International Conference on 3D Vision , pp. 879-888, 11.2020;

[22] C.Lassner, J.Romero, M.Kiefel, F.Bogo, M.J.Black, P.V.Gehler, “Unite the people: Closing the loop between 3d and 2d human representations”, IEEE Conference on Computer Vision and Pattern Recognition, 2017;

[23] K.He, X.Zhang, S.Ren, J.Sun, “Deep residual learning for image recognition”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016;

[24] T.Alldieck, M.Magnor, W.Xu, C.Theobalt, G.Pons-Moll, “Video based reconstruction of 3d people models”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8387–8397, 2018;

[25] I.Makarov,D.Chernyshev. “Real-Time 3D Model Reconstruction and Mapping for Fashion” , 43rd International Conference on Telecommunications and Signal Processing, pp. 133-138, 08.2020;

[26] A.Kanazawal, M.J.Black, D.W.Jacobs, J.Malik1, ”End-to-end Recovery of Human Shape and Pose,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122-7131, 2018.