INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

A Review on Real-Time 3D Reconstruction Methods in Dynamic Scene

Muhammad Nur Affendy Nor'a, Fazliaty Edora Fadzli, Ajune Wanis Ismail
Mixed and Virtual Reality Research Lab, Vicubelab
School of Computing
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia
Email: m.nuraffendy94@gmail.com, ajune@utm.my

Submitted: 14/4/2021. Revised edition: 6/11/2021. Accepted: 8/11/2021. Published online: 16/5/2022
DOI: https://doi.org/10.11113/ijic.v12n1.317

***Abstract*—Advancements made in consumer and readily available RGB-D capturing devices have sparked researcher interest in 3D reconstruction, particularly in dynamic scenes, as well as the quality performance and its speed. The recent advancement in such devices supports the developments of various applications such as teleportation, gaming, volumetric video, and CG films. Real-time 3D reconstruction methods review in a dynamic scene of virtual environment is depicted in this paper. This provides an insight view on how real-time 3D reconstruction beneficial achievement further enables reconstruction systems to be managed in real-time technology such as virtual reality or augmented reality application.**

***Keywords*—3D Reconstruction, 3D Reconstruction Methods, Dynamic Scenes, Real-time 3D Reconstruction**

I. INTRODUCTION

In the grounds of computer graphics and also computer vision, real-time 3D reconstruction as a research topic has been one of the most increasingly interesting areas. 3D reconstruction is a process in which different aspects of the actual visual world, such as object geometry, motion of specific objects while the appearance and texture are observed in the scene, that later are reconstructed in the virtual environment [1]. The capacity to reconstruct any or parts of the real-world elements has opened up new possibilities in the computer graphics also computer vision grounds. Free-viewpoint video can be produced through geometry reconstruction, surface motion, as well as observed appearance while reconstructed kinematic motion is used to create photo realistic animation [2].

Rapid innovations in 3D reconstruction research [3, 4] have made it possible to reliably combine though the fusion method of depth maps from several RGB-D cameras to produce a 3D model of a static scene [5, 6, 7]. Conversely, owing to limitations such as the need for a carefully built capture environment [8, 9] involves high quality equipment and resolution as well as the numerous videos capturing equipment, the task of reconstructing non-rigid scenes is still essentially unsolved. Furthermore, the topic of non-rigid deformation from one shape to another is ill-posed. The usage of single sensor to estimate non-rigid motion is a cumbersome, since more than half of the scene is obscured at any given moment, and because constant movement causes significant frame-to-frame variations, which may contribute to inconsistency in the scene's topological structures, hence increase the difficulty level to create reliable calculations [10]. Fortunately, though devices such as Kinect [11] and Bumblebee [12] have helped overcome acquisition constraints, reconstruction algorithms also need the prior scene to compel the space issue, such as pre-took lighting environments, template of pre-scanned model and intricately embedded skeletons [11]. In this paper we emphasis our review on static and moving camera to reconstruct the dynamic scene. We offer an overview of the various 3D reconstruction methods, with a focus on complex scenes with non-rigid structures, articulated action, or both.

II. Previous Work

The none other well-known previous research work in the 3D reconstruction uses the widely available commodity sensor (Kinect) that incorporates a structured light-based depth sensor. The KinectFusion research work [6] present a system employing only a moving low-cost depth camera and commodity graphics hardware, for a precise real-time mapping of complex and arbitrary indoor scenes in varying illumination conditions. The system uses a simultaneous localization and mapping (SLAM) system to track the environment and do the mapping in real-time. The process consists of (1) surface measurement, (2) surface reconstruction update, (3) surface prediction, and (4) sensor pose estimation.

Then, the KinectFusion in [13] extends the system with a novel interactive reconstruction system where user can dynamically interact with the reconstructed environments through multi-touch interaction. The system able to reconstruct the environment while simultaneously segmenting and tracking foreground objects and the user making it possible to perform 3D reconstruction in a dynamically changing scene.

Apart from working in a rigid dynamically changing scene, researcher has started to investigate the reconstruction of a non-rigid scene in real-time. Research work such as VolumeDeform [14], present a novel dynamic geometric shape for reconstruction using a single RGB-D sensor at real-time rates. The system does not require a template shape to work with the reconstruction as the scene model is build up from scratch during the scanning process. The volumetric representation parameterized the geometry and motion through a distance field of the surface geometry and non-rigid space deformation.

In this paper, we focus on real-time 3D reconstruction method in a dynamic scene. There are several obligatory techniques for reconstructing a 3D dynamic scene due to the design of the camera and object. According to [2], the camera's and world's diverse essence can be divided into four classifications: (1) static camera static object, (2) static camera moving object, (3) moving camera static object, and (4) moving camera moving object as seen in Fig. 1.

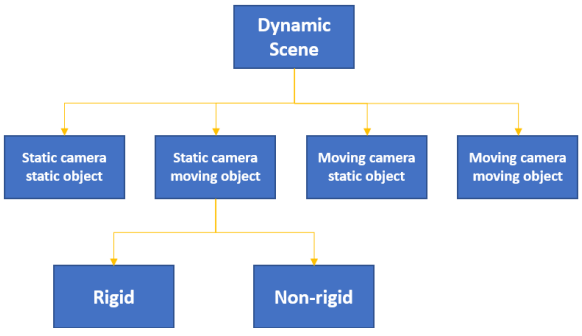

Fig.1. Dynamic scene classification

Further down, the reconstruction in a dynamic scene consists of a rigid and non-rigid 3D reconstruction. A dynamic scene contains both static and dynamic objects. Static objects are those that do not deform, such as a table or a chair, and dynamic objects are those that do deform, such as human interaction and hand movements.

III. Real-Time Non-Rigid 3D Reconstruction in Dynamic Scene

In this paper, we focus on real-time non-rigid 3D reconstruction, which is divided into three sections: (1) general deformation, (2) articulated motion, and (3) human motion capture as illustrated in Fig. 2. Furthermore, these techniques are split further into three groups [2] that consists of: (1) 3D reconstruction of rigid objects, (2) 3D reconstruction of non-rigid objects, (3) 3D reconstruction of articulated motion of the object.

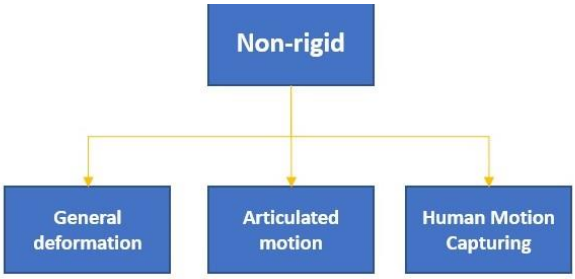

Fig. 2. Non-rigid 3D reconstruction division [2]

In the case of non-rigid forms, as seen in Fig. 3, Zollhöfer [15] proposes a proficient real-time reconstruction system that able to record a range of deformation of non-rigid shapes.

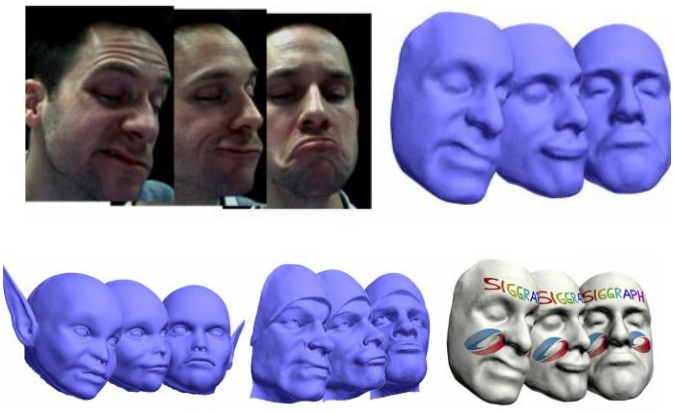

Fig. 3. Non-rigid 3D reconstruction of human face [15

The proposed method is intended for non-rigid reconstructions of single objects at near range. The method consists of two phases where first scanning the desired object while undergoing mostly rigid deformations by using volumetric fusion to automatically extract the triangle mesh.

In order to establish a multi-resolution hierarchy, the mesh is preprocessed before performing real-time non-rigid reconstruction, which results in a deformed mesh at each time stage, by performing the three steps at each frame: (1) rigid registration, (2) non-rigid surface fitting, and (3) detail integration.

Further, the SobolevFusion [16] method focus on reconstruction of free non-rigid motion that is built on Sobolev gradient flow, which allows for a more simple, quicker energy computation while preserving geometric information without over-smoothing. The method manages the topological changes and broad motion, requiring just a few views to create a model with approximation voxel correspondences and color the reconstruction. Fig. 4 below shows the SobolevFusion method reconstruction example acquire from the RGB-D sensor.

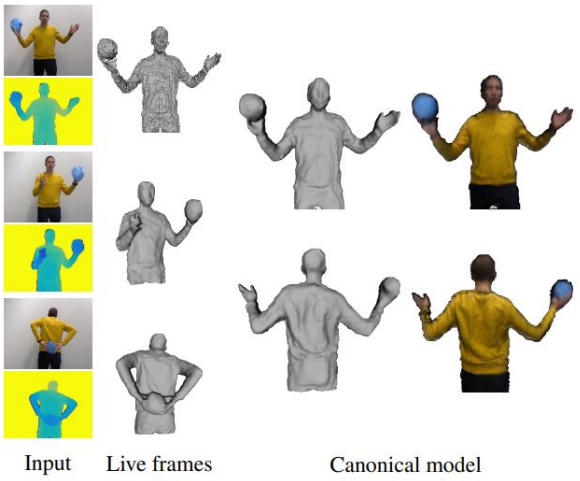

Fig. 4. SobolevFusion reconstruction example [16]

IV. REAL-TIME NON-RIGID 3D RECONSTRUCTION METHODS

Various natural visual world elements, such as intrinsic or observed presence, static geometry and detailed motion are captured and reproduced in dynamic scene reconstruction. A realistic reconstruction rendering for virtual reality scenario as in Holoportation [17] is a perfect example of realistic rendering where the geometry, motion as well as the appearance is reconstructing in real-time. The real-time 3D reconstruction method as in [1], for the non-rigid surface reconstruction in real-time, use dense shading details for a precise and robust surface registration. For casual motion reconstruction, their approach suggested a collection of temporally compatible geometric and photometric correspondences.

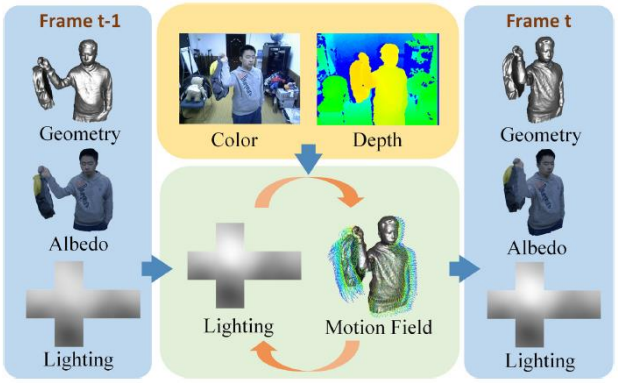

Fig. 5. Joint motion schematic and lighting optimization mechanism [1]

In the proposed method, the method decomposes each frame's photometric data into albedo and low-frequency ambient illumination, for an optimize surface albedo in real-time after several frame fuses using temporally coherent albedo calculation and fusion. The inputs gather from the previous frame are seen in the first column of Fig. 5. The geometry and albedo after warping using the motion field, as well as the illumination that was calculated based on the input color and depth of the current picture, are seen in the last panel.

Another real-time 3D reconstruction method as presented in BodyFusion [18], to increase the reconstruction accuracy of a complex human motion, a skeleton-embedded surface fusion (SSF) technique was introduced. For complex surface fusion, the SSF technique optimizes both the skeleton and the graph-nodes. The technique is claimed to make creating a full-body 3D self-portrait with a single depth camera more simple and real-time.

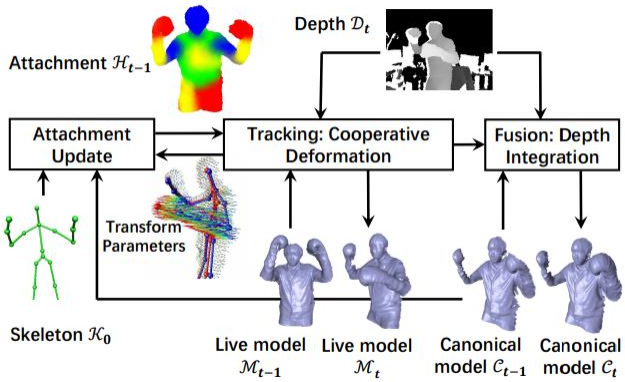

Fig. 6. BodyFusion system pipeline [18]

The illustration as shown in Fig. 6 is the proposed BodyFusion method for real-time 3D reconstruction. The connection update step and the cooperative deformation step, which compensate the deformation of the skeleton as well as the graph-node, are the major discrepancies from the DynamicFusion [19] pipeline.

Apart from the skeleton-based approach in BodyFusion, MaskFusion [20] on the other hand, proposed a method that incorporates Mask-RCNN [21], an effective segmentation algorithm which can forecast labels of object category for 80 object types through its instance level segmentation algorithm based on image. MaskFusion is a real-time functional Simultaneous Localization and Mapping (SLAM) system competent of representing at object level in the scenes. It can recognize, detect, trace, and reconstruct multiple moving rigid objects while precisely segmenting and labeling each case. The method takes advantage of the benefits of integrating Mask-RCNN outputs with a geometry-based segmentation algorithm that produces an object edge map from depth and surface normal clues, allowing it to improve the precision of object borders in object masks.

While in FusionMLS [22], the system incorporates the frame-based method that utilize the moving least square (MLS) reconstruction method that produced a refined set of points on surfaces based on local fitting of point clouds. The FusionMLS proposed system use multiple RGB-D camera setup to capture the real world to be reconstructed. The system is based on a client-server connection where the camera input (client) is sent through the network to a server that decompress the received data and later proceed with reconstruction that was done in the GPU. The reconstruction process consists of two parts, (1) motion estimation and (2) geometric surface estimation, wind-up with a rendering the visuals. Fig. 7 illustrates the FusionMLS system pipeline.

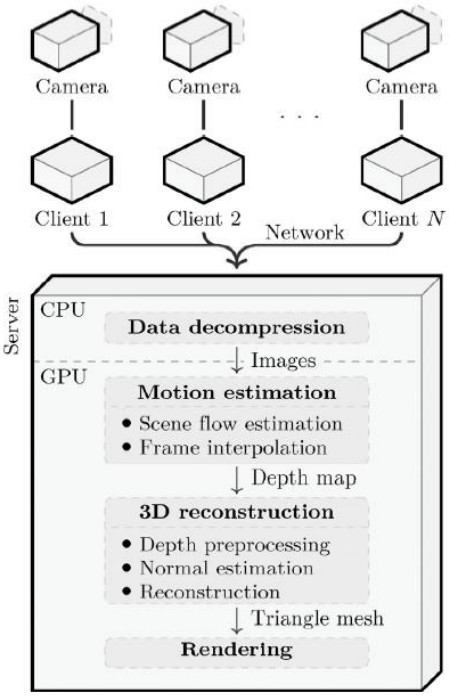

Fig.7. FusionMLS pipeline [22]

Apart from the skeleton-based approach, body shape seems to be a very strong precedent due to its complete and loop closed approach in terms of human performance capture. Hence, the DoubleFusion [23] work take full gain of the shape of the human and its prior motion pose, in their proposed method that reconstruct cloth geometry in real-time also the body form by using the dynamic surface reconstruction method in a single view. In addition, the methods making sure each layer simultaneously gaining advantage from each other. This double-layer surface representation approach reconstructs the model using the outer surface layer, and inner body layer while perform the depth registration. The approach is based on the skinned multi-person linear (SMPL) model [24].

Through the DoubleFusion proposed method, the system allows for real-time simultaneous reconstruction of the outer surface geometry, inner body shape, pose and motion. The method can be achieved by using only one depth camera, and without the need to do any pre-scanning efforts of the desired reconstruction real world model. The method demonstrated that it is capable of significantly improving efficiency in handling rapid movements as well as people dressed casually, all while working in real-time. Fig. 8 shows the results example from the proposed DoubleFusion method of double-layer surface representation approach.

Fig.8. DoubleFusion results example in real-time [23]

Furthermore, the real-time 3D reconstruction proposed method in SimulCap [25] is the enhanced version of DoubleFusion method that based on the double-layer surface representation approach. The SimulCap method main contribution consists of a multi-layer garments and body representation that based on the DoubleFusion [23] and the physics-based performance capture procedure.

The introduced approach aims to create a live free-viewpoint human performance with complex information recorded. Through incorporating fabric simulation into the output capture pipeline, the system will effectively model believable cloth dynamics and its interactions with the human body even in occluded body parts. Furthermore, by defining physical operation of depth fitting, SimulCap able to produce consistent results of cloth tracking with the depth observation while remaining physically constrained. Fig. 9 shows the results in real-time of the 3D reconstruction method proposed in SimulCap.

Fig. 9. SimulCap results example in real-time [25]

In a real-time 3D reconstruction method, tedious self-scanning approach to initialize the reconstruction process can be a major constraint. Due to this issue, RobustFusion [26] method proposed a novel optimization pipeline that only uses the front-view input and later combine with the data-driven volumetric fusion occupancy representation. The robust human volumetric capture process presented in RobustFusion incorporate a variety of visual cues that driven by data in a monocular setting devoid of the use of a pre-scanned template. RobustFusion method able to effectively capture robust performance through its arrangement of human posture, form, parsing priors, that able to handle difficult human movements with the capacity to reinitialize. The RobustFusion pipeline, as seen in Fig. 10, involves of a model completion stage and a stable performance capture stage for live 4D results, assuming monocular RGBD input with numerous data-driven human visual priors.

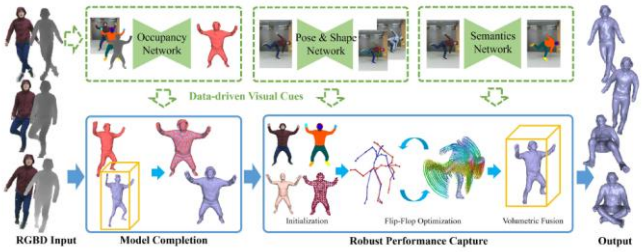

Fig. 10. RobustFusion pipeline [26]

The Function4D on the other hand, proposed a volumetric capture pipeline that consist of 2 steps: (1) Dynamic Sliding Fusion (DSF) that fuse neighboring frames from the gathered frames from multiple camera input to produce a noise free and temporal continuous result, and (2) Deep Implicit Surface Reconstruction that re-render the DSF back into the original viewpoints. Then proceed with the implicit function to generate a complete and detail reconstruction output. Fig. 11 illustrates the system pipeline.

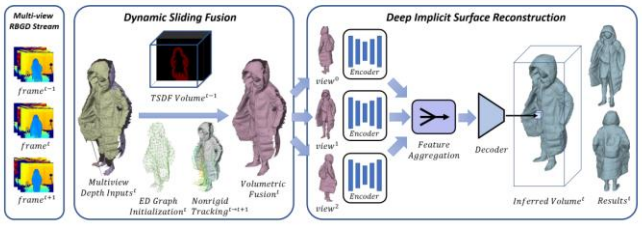

Fig. 11. Function4D system process [27]

V. Discussion

In performing a real-time 3D reconstruction, it is crucial to take into account the appropriate approach in order to produce a realistic look of the reconstructed model. This is inclusive with the number of camera input in the system setup as higher number of cameras may produce a high-resolution reconstructed model but may be highly cost. Table 1 presents the research works in 3D reconstruction realm with its proposed camera setup that incorporates either single or multiple, and static or moving camera placement.

TABLE I. 3D Reconstruction Camera Setup

Year	Research Works	Camera Setup			
		Single-camera	Multiple-camera	Static	Moving
2017	Real-Time Geometry, Albedo, and Motion Reconstruction using a Single RGB-D Camera [1]	✔		✔	
2017	BodyFusion [18]	✔		✔	
2018	SobolevFusion [16]	✔		✔	
2018	MaskFusion [20]	✔			✔
2018	FusionMLS [22]		✔	✔	
2018	DoubleFusion [23]	✔		✔	
2019	SimulCap [25]	✔		✔	
2020	RobustFusion [26]	✔		✔	
2021	Function4D [27]		✔	✔	

As we can see from Table 1, only FusionMLS [22] and Function4D [27] use a multiple camera setup while only MaskFusion [20] allow for camera movement during reconstruction. This is due to their proposed method that require multiple camera setup and use the client-server architecture as in FusionMLS [22], and to highlight the capability of their proposed method MaskFusion [20] with Mask-RCNN [21] adoption that capable of an instance level segmentation algorithm in a dynamic scene either moving objects or moving camera. The other research works presented in the Table 1 use the same single camera setup with a static camera arrangement as it to comply and suited with their proposed method that works with the mentioned camera setup.

As for research work in [1] that focus on albedo, usage of low-frequency environmental lighting of the reconstructed model, BodyFusion [18] that incorporates skeleton-embedded surface fusion that only need single camera setup to work, the SobolevFusion [16] method that based on Sobolev gradient flow, which allows for a more simple, quicker energy computation while preserving geometric information without over-smoothing using single camera approach, and DoubleFusion [23] that take full gain of shape of human and its prior posture, in their proposed method of a

single-camera setup and real-time dynamic surface reconstruction system that simultaneously reconstructs general cloth geometry and inner body shape. While for SimulCap [23] that enhance the work in DoubleFusion with additional physics-based performance capture procedure, and RobustFusion [26] that based on data-driven that use only the front-view input of a single camera rig, with a model completion structure that used for high-resolution initial model creation and completion.

Through the data presented in Table 1 shows that most of the researcher adopt the single camera setup with a static position of the camera placement. Researcher that adopts such method typically will implement the reconstruction method that based on prior information such as template or skeleton based, visual data-driven and the usage of machine learning [1, 18, 16, 20, 23, 25, 26]. While as for the researcher that adopt the multiple camera setup imposed their reconstruction method directly on the received frames from multiple camera input [22, 27]. Both of the camera setup and placement poses its own advantages as for single camera setup may be low in cost compared to multiple camera setup but may require an expensive computational process in retrieving or estimating the model surface reconstruction. Another integral downside of single camera setup is these methods are susceptible to tracking failure in hidden areas of the scene that may be due to occlusion [27].

In conclusion, in this paper, we summarize the methods used to solve the issues of real-time 3D reconstruction of non-rigid objects in a dynamic environment. Despite the fact that some proposed method and enhancement have solved problems in real-time 3D modeling of non-rigid structures, there are still many remaining challenges to be addressed such as methods to achieve high resolution and preserved detailed reconstructed 3D model in real-time and also in the area of a topological aware surface reconstruction of the reconstructed model that may suffer due to open-close movement in the dynamic scene [28].

ACKNOWLEDGMENT

We would like to express our appreciation to Mixed and Virtual Reality Laboratory (mivielab) in Vicubelab at Universiti Teknologi Malaysia (UTM). This work was funded by UTM-GUP Funding Research Grants Scheme (Q.J130000.2628.14J85).

REFERENCES

[1] Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., & Liu, Y. (2017). Real-time Geometry, Albedo, and Motion Reconstruction Using a Single Rgb-d Camera. *ACM Transactions on Graphics (ToG)*, 36(4), 1.

[2] Ingale, A. K. (2021). Real-time 3D Reconstruction Techniques Applied in Dynamic Scenes: A Systematic Literature Review. *Computer Science Review*, 39, 100338.

[3] Divya Udayan, J., & Kim, H. (2016). Constrained Procedural Modeling of Real Buildings from Single Facade Layout. *International Journal of Computer Vision and Signal Processing*, 6(1).

[4] Udayan, J. D. (2016). An Analysis of Reconstruction Algorithms Applied to 3d Building Modeling. *Indian Journal of Science and Technology*, 9, 33.

[5] Curless, B., & Levoy, M. (1996). A Volumetric Method for Building Complex Models from Range Images. Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 303-312.

[6] Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., ... & Fitzgibbon, A. (2011). Kinectfusion: Real-time Dense Surface Mapping and Tracking. *2011 10th IEEE International Symposium on Mixed and Augmented Reality, IEEE, 127-136.*

[7] Zach, C. (2008). Fast and High Quality Fusion of Depth Maps. *Proceedings of the International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 1(2).

[8] De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H. P., & Thrun, S. (2008). Performance Capture from Sparse Multi-view Video. *ACM SIGGRAPH 2008 papers*, 1-10.

[9] Vlasic, D., Peers, P., Baran, I., Debevec, P., Popović, J., Rusinkiewicz, S., & Matusik, W. (2009). Dynamic Shape Capture Using Multi-view Photometric Stereo. *ACM SIGGRAPH Asia 2009 papers*, 1-11.

[10] Li, H., Adams, B., Guibas, L. J., & Pauly, M. (2009). Robust Single-view Geometry and Motion Reconstruction. *ACM Transactions on Graphics (ToG)*, 28(5), 1-10.

[11] Sarbolandi, H., Lefloch, D., & Kolb, A. (2015). Kinect Range Sensing: Structured-light versus Time-of-Flight Kinect. *Computer Vision and Image Understanding*, 139, 1-20.

[12] Marani, R., Renò, V., Nitti, M., D'Orazio, T., & Stella, E. (2015). A Compact 3D Omnidirectional Range Sensor of High Resolution For Robust Reconstruction of Environments. *Sensors*, 15(2), 2283-2308.

[13] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., ... & Fitzgibbon, A. (2011). KinectFusion: Real-time 3D Reconstruction and Interaction using a Moving Depth Camera. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 559-568.

[14] Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., & Stamminger, M. (2016). Volumedeform: Real-time Volumetric Non-rigid Reconstruction. *European Conference on Computer Vision*, Springer, Cham, 362-379.

[15] Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., ... & Stamminger, M. (2014). Real-time Non-rigid Reconstruction Using an RGB-D Camera. *ACM Transactions on Graphics (ToG)*, 33(4), 1-12.

[16] Slavcheva, M., Baust, M., & Ilic, S. (2018). Sobolevfusion: 3d Reconstruction of Scenes Undergoing Free Non-rigid motion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2646-2655.

[17] Fadzli, F. E., Ismail, A. W., Aladin, M. Y. F., & Othman, N. Z. S. (2020). A Review of Mixed Reality Telepresence. *IOP Conference Series: Materials Science and Engineering*, 864(1), 012081. IOP Publishing.

[18] Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., ... & Liu, Y. (2017). Bodyfusion: Real-time Capture of Human Motion and Surface Geometry Using a Single Depth Camera. *Proceedings of the IEEE International Conference on Computer Vision*, 910-919.

[19] Newcombe, R. A., Fox, D., & Seitz, S. M. (2015). Dynamicfusion: Reconstruction and Tracking of Non-rigid scenes in Real-time. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 343-352.

[20] Runz, M., Buffier, M., & Agapito, L. (2018). Maskfusion: Real-time Recognition, Tracking and Reconstruction of Multiple Moving Objects. 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 10-20.

[21] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2961-2969.

[22] Meerits, S., Thomas, D., Nozick, V., & Saito, H. (2018). Fusionmls: Highly Dynamic 3d Reconstruction with Consumer-grade rgb-d Cameras. *Computational Visual Media*, 4(4), 287-303.

[23] Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., ... & Liu, Y. (2018). Doublefusion: Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7287-7296.

[24] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. 2015. SMPL: A Skinned Multi-person Linear Model. *ACM Transactions on Graphics (TOG)*, 34(6), 1-16.

[25] Yu, T., Zheng, Z., Zhong, Y., Zhao, J., Dai, Q., Pons-Moll, G., & Liu, Y. (2019). Simulcap: Single-view Human Performance Capture with Cloth Simulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5504-5514.

[26] Su, Z., Xu, L., Zheng, Z., Yu, T., & Liu, Y. (2020). Robust Fusion: Human Volumetric Capture with Data-Driven Visual Cues Using a RGBD Camera. ECCV.

[27] Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., & Liu, Y. (2021). Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5746-5756.

[28] Li, C., & Guo, X. (2020, August). Topology-Change-Aware Volumetric Fusion for Dynamic Scene Reconstruction. In *European Conference on Computer Vision*, Springer, Cham, 258-274.