Fusion4D: Real-time Performance Capture of Challenging Scenes

Anna Baumeister

Department of Informatics - Technische Universität München

Abstract

Reconstruction of challenging performances is a task traditionally confined to offline systems, with most state-of-the art online reconstruction methods limited to scenes with small frame-to-frame motion and unchanging topology. With Fusion4D, Dou et al. present a new processing pipeline that provides real-time reconstruction of challenging scenes in a multi-view setup. During the live reconstruction, a reference model is maintained and refined over time using a new concept called key volumes, combining the advantages of the reference-to-frame and frame-to-frame model. For each frame the reference model is deformed non-rigidly to fit the current input data. The quality of this alignment is modelled by an energy equation which is optimized fully in parallel on the GPU. Finally, the current data volume and warped reference volume are fused to achieve a high quality output model with very little noise. The approach is shown to be robust to many complex topology changes, occlusions as well as large motion and is the first of its kind for real-time reconstruction systems for multi-view performance capture.

1 Introduction

A majority of real-time reconstruction systems focus on capturing static scenes or objects undergoing rigid movement due to the high complexity and time requirements of modelling free non-rigid deformations. With Fusion4D[5], Dou et al. present one of the first real-time reconstruction systems for multi-view performance capture as part of their work on Holoportation [7].

1.1 Related Work

There have already been several publications that achieve high quality volumetric reconstruction in real time but lack robustness to challenging scenes. Zollhöfer et al. demonstrated live reconstruction of diverse objects under non-rigid deformation with a single RGB-D camera [10]. This approach requires the acquisition of a template under rigid deformations prior to the live reconstruction. Due to the template defining the topology of the scene, this approach is not robust to topology changes. Fast motion from frame to frame is also problematic since this this method relies heavily on closest point correspondence matching. In this approach the reference model, once acquired, does not change based on new input data and is only used to explain the current scan. DynamicFusion achieves live reconstruction without the need for prior template acquisition by employing non-rigid volumetric fusion [6]. In this method, a reference volume of the scene is maintained throughout the reconstruction and refined with every new input scan. This way, the template can be improved over time and parts of the model that have previously been occluded can be filled in as new data arrives. However, this system is still not robust to large frame-to-frame motion and topology changes since the reference model can become out-of-date with the current scene.

2 The Fusion4D Pipeline

The Fusion4D Pipeline is illustrated in Figure 1. It can be broadly separated into three steps that are repeated for each frame:

- Data acquisition,
- Non-rigid alignment,
- Volumetric fusion and blending.

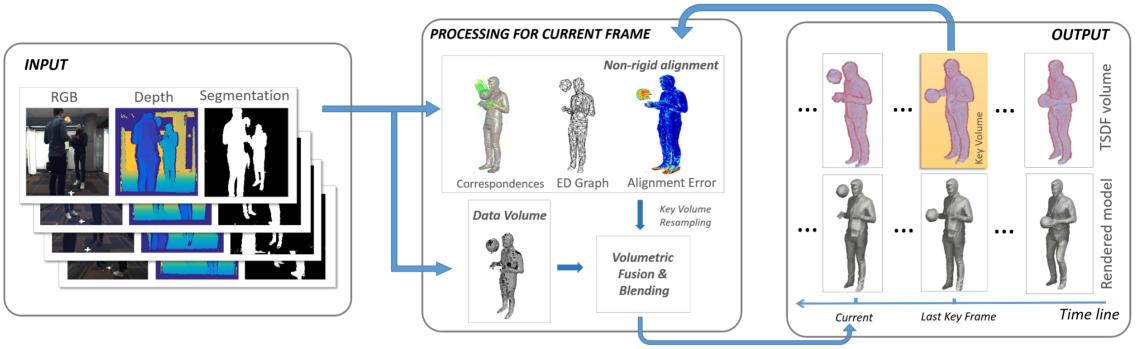

Figure 1: The Fusion4D pipeline. First, an RGB image, a depth map and a segmentation map are obtained for each camera. A *key volume* representing the scene is warped to fit the current input data. Furthermore, a *data volume* is sampled from the input depth maps. Finally, the key volume is refined using the current depth information and the final output volume is obtained by fusing the data volume and warped reference volume.

2.1 Data Acquisition

The performance is captured by a setup of eight custom camera rigs, each with an active stereo setup (two NIR sensors and a structured light source) and one RGB camera. A depth map for each camera is obtained in real time using the PatchMatch Stereo algorithm introduced by Bleyer et al.[2]. For each frame, a segmentation mask is estimated using real-time foreground-background segmentation with a simple background model. This segmentation mask represents the region of interest in the depth map and makes it possible to capture performances in natural settings without a greenscreen.

2.2 Non-rigid Alignment

Throughout the reconstruction, a reference model is maintained to store information about the scene. During the non-rigid alignment step this model is warped to match the input data. After the deformation parameters are obtained, the warped reference model the current data volume are fused and blended to obtain a high quality output volume.

2.2.1 Choice of Reference Frame

The choice of which reference model to use as a template for the current frame is crucial in being able to handle large frame-to-frame movements and topology changes. One intuitive approach that is also employed in the DynamicFusion pipeline is to use a single reference model that is acquired at the first data frame and maintained throughout the reconstruction. For each frame, this reference volume is warped to describe the current data volume. The current data volume is then fused back into the reference volume using the estimated deformation parameters, improving the quality of the reference volume over time. A problem with this approach is that a change in topology breaks the system since the reference volume no longer describes the scene well and the non-rigid matching can find no way to align the model to the new data.
Another method is to always use the volume obtained in the last frame as reference volume for the next frame. This way, topology changes can be correctly modelled since a topology change in the scene will be incorporated into the reference model after just one frame. Furthermore, the reference model is always similar to the current data volume, making the correspondence evaluation between two consecutive frames easier and the non-rigid alignment converge faster. However, repeated resampling of the reference volume with every frame leads to increased geometric blur.
Dou et al. propose a new concept for choosing a reference frame called *key volumes*, which is related to key frames or anchor frames[1] and combines the advantages of both aforementioned methods. Conceptually this approach is similar to the reference-to-frame method used for DynamicFusion in that it fixes a reference volume at the first frame and models subsequent frames by deforming

this model. This key volume is then refined over time by fusing with new data volumes. However, to be able to correctly model the scene even after a topology change has occurred, the key volume is periodically reset every ten frames. The new key volume is then obtained by fusing the old key volume and the new current data volume. Thus, a change in topology is incorporated into the reference model after at most ten frames which is sufficient even for fast movement, while the quality of the reference model improves steadily as new input data is fused into it.

2.2.2 Non-rigid Matching to the Data Frame

For every frame, the key volume is warped to match the current input data consisting of a depth image and segmentation map containing the region of interest. To describe the deformation in an efficient manner, the reference volume is subsampled uniformly at random to obtain a graph of *embedded deformation* or *ED* nodes as described in [8]. The global non-rigid deformation of the reference model can then be described by a sum of local affine transformations of the ED nodes. The deformation of any surface vertex of the model can then be described using *linear blend skinning* of all ED nodes that influence this vertex. The quality of the alignment between the warped reference model and the input data is described by the energy formulation

$$\boldsymbol{E}(\boldsymbol{G}) = \boldsymbol{\lambda}_{data}\boldsymbol{E}_{data} + \boldsymbol{\lambda}_{rot}\boldsymbol{E}_{rot} + \boldsymbol{\lambda}_{hull}\boldsymbol{E}_{hull} + \boldsymbol{\lambda}_{smooth}\boldsymbol{E}_{smooth} + \boldsymbol{\lambda}_{corr}\boldsymbol{E}_{corr}. \tag{1}$$

The most important term in this formulation is the data term. It describes the quality of the alignment of the reference frame to the current data frame using a projective point-to-plane approximation. However, using only $\boldsymbol{E}_{data}$ to assess the quality of the alignment can easily lead to unreasonable deformations. Thus, further regularization terms are introduced that enforce a more natural looking deformation.
The rotational term $\boldsymbol{E}_{rot}$ rewards affine transformations of ED nodes that are close to a rotation. Thus, it encourages the local deformations to be close to a rigid transform.
The smoothing term $\boldsymbol{E}_{smooth}$ encourages the affine transformations of two neighbouring nodes on the ED graph to be similar, leading to a smoother surface in the reconstructed model.
Another regularization term is $\boldsymbol{E}_{hull}$, which encourages points of the deformed model to lie within the visual hull. The visual hull is obtained by intersecting the projections into free space of the object silhouette observed by each camera. Intuitively, the visual hull forms a bounding box around the object and gives a hard constraint on where all deformed data points must lie. Space outside the visual hull was observed as free space by multiple cameras and should not be occupied by a deformed reference model point.
The final term in the energy function $\boldsymbol{E}_{corr}$ alone is based on the RGB input data. Using an extension of the Global Patch Collider algorithm introduced by Wang et al. [9] correspondences between two consecutive input frames are determined. Reference model points are then encouraged to transform to their 3D correspondences.
The factors $\boldsymbol{\lambda}$ govern the strength of each parameter and are chosen empirically to achieve the best reconstruction result.

2.2.3 Optimization

The goal of each non-rigid matching step is to find a set of parameters $\boldsymbol{G}$ that transforms the ED nodes in such a way that the energy function is minimized. Since $\boldsymbol{E}$ is just a sum of square residuals, it can be rewritten as:

$$\boldsymbol{E}(\boldsymbol{X}) = \boldsymbol{f}(\boldsymbol{x})^T\boldsymbol{f}(\boldsymbol{x}) \tag{2}$$

where $\boldsymbol{f}(\boldsymbol{x})$ is a vector $\in \mathbb{R}^D$ containing the unsquared terms of $\boldsymbol{E}$ (i.e. the residuals). Thus, minimizing $\boldsymbol{E}(\boldsymbol{X})$ can be seen as a standard linear least squares problem that can be solved by Gauss-Newton-like methods like the Levenberg-Marquardt(LM) algorithm. In each iteration of the LM algorithm, the new step direction $\boldsymbol{h}$ is obtained by solving

$$(\boldsymbol{J}^T\boldsymbol{J} + \mu\boldsymbol{I})\boldsymbol{h} = -\boldsymbol{J}^T\boldsymbol{f}(\boldsymbol{x}), \tag{3}$$

Where $\boldsymbol{J}$ is the Jacobian matrix and μ a damping factor that controls how aggressive the next step of the solver will be. Calculating $\boldsymbol{h}$ means solving a system of linear equations as defined by

Equation (3). DynamicFusion obtains a solution by using a sparse Cholesky decomposition which requires explicit evaluation of $\boldsymbol{J}^T\boldsymbol{J}$. Since evaluating $\boldsymbol{J}^T\boldsymbol{J}$ directly is not feasible in a real-time system, it is approximated as a block diagonal matrix, which allows the system to run in real time but severely affects the quality of the reconstruction when large frame-to-frame motion occurs. In order to not impair the fidelity of the reconstruction, Dou et al. instead approximate a solution to Equation (3) by using ten steps of *preconditioned gradient descent* using the block diagonal approximation to $\boldsymbol{J}^T\boldsymbol{J}$ as a preconditioner. This approximation gives results that are comparable to a full analytic solve and still runs in real time.

2.3 Volumetric Fusion and Blending

In a final step, the data volume is fused and blended with the warped reference volume to achieve a high quality output volume. Additionally, the reference volume for the next input frame is generated by fusing the warped key volume, the last reference volume and the current data volume.

2.3.1 Fusion at the Data Frame

In this step, the non-rigid alignment of the reference frame to the data frame is used to increase the quality of the data volume and output a high quality volumetric model of the scene. Volumes are represented as a two-level hierarchy according to [3]. For each frame, a data volume is estimated for the current depth- and segmentation map. The current reference volume is warped to fit this data volume using the deformation parameters estimated in the non-rigid matching phase. Since the quality of the data volume should only be improved by fusion with the warped reference volume, selective volumetric fusion is employed. Before fusing a warped voxel, two tests are performed and the voxel is rejected if it fails either:

- **Voxel collision**: Reference voxels that contribute to different surface areas might collide after warping, for example in the case of clapping hands. in this case, the TSDF representation of the two surfaces might average out, leading to holes or other inconsistencies in the fused mesh.
- **Voxel misalignment**: Only reference voxels whose corresponding ED nodes have a small alignment error should be accepted for volumetric fusion in order to not decrease the quality of the output volume.

3 Results

Fusion4D is shown to give visually compelling reconstruction results compared to other state-of-the-art online systems. It is robust to many complex topology changes (e.g. a performer taking off a jacket, catching a ball, clapping hands), fast frame to frame motion and occlusions while still maintaining a high reconstruction quality at 31 frames per second. Figure 2 shows reconstruction of a live performance that has both large movements and closed-to-open topology changes as the hands move from the hips over the head of the performer. Fine details like the folds on the skirt are reconstructed faithfully even when large motion is present.

Figure 2: Real-time reconstruction of a challenging scene using the Fusion4D pipeline. Fine details like the hair and dress of the performer are reconstructed faithfully. Furthermore, the large topology change caused by the performer moving her hands does not break the reconstruction.

A more quantitative comparison is given with the 'Breakdancers' dataset by Collet et al. which exhibits extremely high frame-to-frame motion, topology changes and occlusions [4]. The Fusion4D pipeline is shown to give robust reconstruction results even while other state-of-the-art online systems like DynamicFusion fail. Furthermore, it gives results which are comparable to a high quality offline system that has processing times as high as 30 seconds for each input frame.
Even though the Fusion4D pipeline has exhibited compelling live reconstruction results, it is not without limitations. If tracking is lost at any point during the reconstruction (for example due to the non-rigid matching not converging within the time limit of 33ms) the system falls back to the live fused data. During the following frames the output volume can look noisy while new information is first being fused in. An error in the estimated segmentation map, e.g. due to missing depth information can lead to an inaccurate estimation of the visual hull, causing noise or holes to be fused into the model. Furthermore, errors in the non-rigid alignment can at times lead to oversmoothing of the model.

4 Conclusion

With Fusion4D, Dou et al. present the first real-time non-rigid reconstruction pipeline for multi-view performance capture. Given noisy input from multiple cameras it produces a temporally consistent 3D model that is comparable to high quality offline reconstruction methods. The approach is shown to robustly handle large frame-to-frame motion as well as topology changes using a novel correspondence algorithm and real-time optimizer. Even though the system still has some limitations, it offers new possibilities for capturing and broadcasting live performances like sporting events in 3D.

References

[1] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)*, volume 30, page 75. ACM, 2011.

[2] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

[3] Jiawen Chen, Dennis Bautembach, and Shahram Izadi. Scalable real-time volumetric surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(4):113, 2013.

[4] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):69, 2015.

[5] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016.

[6] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

[7] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754. ACM, 2016.

[8] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)*, 26(3):80, 2007.

[9] Shenlong Wang, Sean Ryan Fanello, Christoph Rhemann, Shahram Izadi, and Pushmeet Kohli. The global patch collider. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 127–135, 2016.

[10] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33(4):156, 2014.