STAG: Smart Tools and Applications in Graphics (2022)
D. Allegra, C. E. Catalano, G. Cherchi, and R. Scateni (Editors)

Poster

Deep tracking for robust real-time object scanning

M. Lombardi[1], M. Savardi[1,2] and A. Signoroni[1,2]

[1]University of Brescia, Department of Information Engineering, Italy
[2]University of Brescia, Dept. of Medical and Surgical specialties, Radiological Sciences and Public Health, Italy

Abstract

Nowadays, a high-fidelity 3d model representation can be obtained easily by means of handheld optical scanners, which offer a good level of reconstruction quality, portability, and low latency in scan-to-data. However, it is well known that the tracking process can be critical for such devices: sub-optimal lighting conditions, smooth surfaces in the scene, or occluded views and repetitive patterns are all sources of error. In this scenario, recent disruptive technologies such as sparse convolutional neural networks have been tailored to address common problems in 3D vision and analysis. Our work aims to integrate the most promising solutions into an operating framework which can then be used to achieve compelling results in 3D real-time reconstruction. Several scenes from a dataset containing dense views of objects are tested using our proposed pipeline and compared with the current state-of-the-art of online reconstruction.

CCS Concepts

• ***Computing methodologies*** → *Reconstruction; 3D imaging; Tracking; Artificial intelligence;* • ***Hardware*** → *Emerging tools and methodologies;*

1. Introduction

The 3D reconstruction with handheld devices without the aid of targets (i.e. solely based on geometric tracking) presents various solutions in the literature. The pioneer in the sector was the work Kinectfusion [NIH*11, IKH*11, INK*11, NLD11] that introduced some key elements to manage the alignment of a sequence of frames, namely the use of a "local" algorithm such as ICP [BM92, CM91, RL01] in a frame-to-model approach. In this context, the scene is built by integrating the various views within a volume [CL96] from which a snapshot is periodically extracted via ray-casting to represent the model to which the frames will be aligned. In a subsequent work [GISC13], Glocker *et al.* tried to implement a camera relocation after tracking loss by introducing the use of key-frames encoded by *randomized ferns* exploiting depth and appearance information. Niessner *et al.* [NZIS13] focused instead on reducing memory consumption through voxel hashing of a sparse volume. More recently, BundleFusion system [DNZ*17] introduced some key improvements such as 1) the hierarchical approach to reconstruction, dividing the sequence of frames into chunks, 2) a coarse-to-fine alignment based on 2D feature matching (SIFT [Low04]), to strengthen the tracking and, 3) the dynamic update of the model, through the volumetric reintegration process. In ElasticFusion [WSMG*16] and VolumeDeform [IZN*16] the focus was on real-time reconstruction of non-rigid objects, while Xiang *et al.* [XJZ*21], Han *et al.* [HGZ*22] and Prisacariu *et al.* [PKMR15] improved the visual rendering of textures and reduced computational complexity to bring the reconstruction on mobile systems, and to create applications more oriented towards augmented and virtual reality. Nevertheless, geometric accuracy remains of vital importance for those metrological-like applications that require a reliable level of detail and are more interested in the 3D model quality. Examples of such applications are reverse engineering, manufacturing, robotics, and also orthotics [VBS18].
In this context, representation learning approaches gave a breakthrough to 3D analysis application field. Deep architectures, indeed, can be learned to infer complex patterns and to establish spatial relationships which are harder to define (and generalize also) with handcrafted solutions. In order to consume point clouds, the earlier attempts involved the use of dense solutions. Specifically, PointNet [CSKG17, QYSG17] constructed Fully Connected networks by means of Multi Layer Perceptrons (MLPs) to address the classification and segmentation tasks with however memory consumption limitations and low spatial coherence handling. Recently, Choy et al. [CGS19] proposed an alternative by leveraging the generalization of convolution via sparse operations on sparse tensors. This solution seems more robust and does not suffer from the lack of locality.
Following a pilot benchmark study on state-of-the-art representation learning techniques for 3D view alignment [LSS20] we propose a fully geometrical pipeline for real-time object model reconstruction. This solution grounds an implementation of the method proposed in [LSS21a] that introduces a safeguard module able to detect failed incremental alignment attempts (ICP-driven) and to activate a deep feature learning-driven coarse alignment module to help alignment tracking recovery. In particular, we introduce a hier-

DOI: 10.2312/stag.20221262

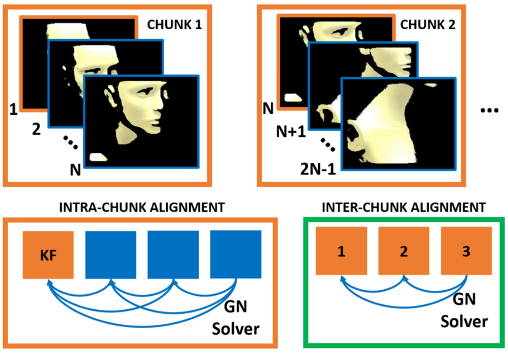

Figure 1: *Intra-chunk and inter-chunk pose alignment for BundleFusion. Multi-view registration exploits Gauss Newton (GN) solver.*

archical pose optimization process different from that proposed in BundleFusion, and better suited to improve the global consistency of the model when using Deep Learning-based solutions. We test our system on the recently proposed DenseMatch dataset [LSS21b] for quantitative analysis of object scanning assessment.

2. Pose optimization via hierarchical-based approach

In order to seek the global consistency in a real-time working environment, we are initially inspired by what proposed in BundleFusion (BF) [DNZ*17] adopting a hierarchical subdivision of the frames sequence, albeit with some differences. BF seeks the global model consistency by proposing a hierarchical subdivision of the frames composing the reconstructing sequence. The new structure of the sequence has 2 levels in its hierarchy: the local and the global level. In Fig.1 we see the main components. Indeed, at the first level of the hierarchy we perform the *local* fine and dense registration, which aligns the current frame against the most recent key-frame. This is very similar to what KinectFusion also does. During this process, BF also extracts 2D features via SIFT that are used for a coarse pre-alignment of the frame, and then it investigates a multitude of key-frames to find the closer one. Such an approach is similar also to what can be found in [GISC13]. However, 2D feature based methods both need to work with heavy down-sampled images and rely solely on 2D similarities, which can be extremely tricky when the target of the reconstruction is a small-size object with a smooth or repetitive texture. Our experiments showed that this is not only an overkill for our purpose, but can also mislead the alignment badly. For the same reason, we also found unnecessary to perform the first level optimization in which BF attempts to refine a chunk o frames by performing an all-vs-all realignment. On the contrary we found extremely helpful to perform the refinement at the second level of the hierarchy, *i.e.* when the optimization is performed globally. In this case, frames are clustered into N chunks according to their temporal consistency, then each chunk is treated as an independent pointcloud. Triggered by the safeguard condition from [LSS21a] we leverage the coarse registration module (which is powered by FCGF [CPK19] and DGR [CDK20]) to pre-align the data on-the-fly (see Fig. 2). We then create a pose graph as suggested by Choi [CZK15], so that we can finally optimize the graph to infer the updated poses of each chunk. After pose optimization, real-time constraints require computationally efficient model updates possibly avoiding to recompute the integration of all the frames. We are currently developing solutions for this by adapting to our context also the dynamic model update approach.

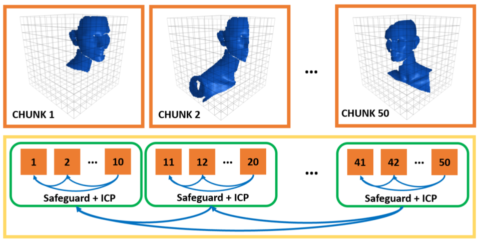

Figure 2: *Our inter-chunk refinement. Pairwise registrations are performed using the safeguard module and ICP. The multi-view is exploited via graph optimization [CZK15].*

3. Results

Our computational pipeline has been implemented in Python with the use of Open3D [ZPK18] as the open-source library to manage the 3D processing, whereas we exploited the MinkowskiEngine [CGS19] and PyTorch the for sparse neural network implementation. To test our pipeline we decided to consider some scenes from the DenseMatch [LSS21b] dataset.With our tests we show the potential of our pipeline and highlight what has already been anticipated: the current state-of-the-art, namely BF, focuses too much on the 2D appearance to perform the coarse pre-alignment during the reconstruction, which results being misleading in our working scenario. When reconstructing scenes on the on DenseMatch dataset [LSS21b], our method reach an average number of correct matches of 98.0% over a total of 6026 frames, outclassing BF that only reaches 19.9% on the same data. Finally, our pipeline ends up behaving better not only in terms of robustness by increasing the total number of aligned frames, but also in terms of the quality of the reconstruction, as evidenced by the results shown in figure 3. Eventually, a GPU implementation of the pipeline guarantees alignment times compatible to handheld scanners operating at about 20 fps.

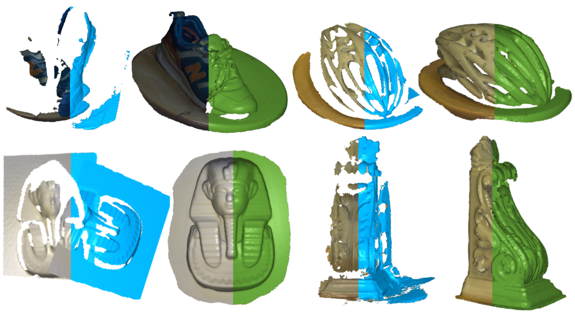

Figure 3: *Qualitative examples of reconstructed scenes from DenseMatch dataset [LSS21b]. In blue: results from BundleFusion. In green: results from our pipeline (texture and mesh are shown).*

References

[BM92] Besl P. J., McKay N. D.: A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14*, 2 (2 1992), 239–256. doi:10.1109/34.121791. 1

[CDK20] Choy C., Dong W., Koltun V.: Deep global registration. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 2511–2520. doi:10.1109/CVPR42600.2020.00259. 2

[CGS19] Choy C., Gwak J., Savarese S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, 6 2019), IEEE Computer Society, pp. 3070–3079. doi:10.1109/CVPR.2019.00319. 1, 2

[CL96] Curless B., Levoy M.: A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1996), SIGGRAPH '96, ACM, pp. 303–312. doi:10.1145/237170.237269. 1

[CM91] Chen Y., Medioni G.: Object modeling by registration of multiple range images. In *Proceedings. 1991 IEEE International Conference on Robotics and Automation* (4 1991), pp. 2724–2729 vol.3. doi:10.1109/ROBOT.1991.132043. 1

[CPK19] Choy C., Park J., Koltun V.: Fully convolutional geometric features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Los Alamitos, CA, USA, 11 2019), IEEE Computer Society, pp. 8957–8965. doi:10.1109/ICCV.2019.00905. 2

[CSKG17] Charles R. Q., Su H., Kaichun M., Guibas L. J.: Pointnet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition* (7 2017), pp. 77–85. doi:10.1109/CVPR.2017.16. 1

[CZK15] Choi S., Zhou Q.-Y., Koltun V.: Robust reconstruction of indoor scenes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 5556–5565. doi:10.1109/CVPR.2015.7299195. 2

[DNZ*17] Dai A., Niessner M., Zollhöfer M., Izadi S., Theobalt C.: Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph. 36*, 3 (May 2017). doi:10.1145/3054739. 1, 2

[GISC13] Glocker B., Izadi S., Shotton J., Criminisi A.: Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2013), pp. 173–179. doi:10.1109/ISMAR.2013.6671777. 1, 2

[HGZ*22] Han L., Gu S., Zhong D., Quan S., Fang L.: Real-time globally consistent dense 3d reconstruction with online texturing. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 3 (2022), 1519–1533. doi:10.1109/TPAMI.2020.3021023. 1

[IKH*11] Izadi S., Kim D., Hilliges O., Molyneaux D., Newcombe R., Kohli P., Shotton J., Hodges S., Freeman D., Davison A., Fitzgibbon A.: 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* (New York, NY, USA, 2011), Association for Computing Machinery, pp. 559–568. URL: https://doi.org/10.1145/2047196.2047270, doi:10.1145/2047196.2047270. 1

[INK*11] Izadi S., Newcombe R. A., Kim D., Hilliges O., Molyneaux D., Hodges S., Kohli P., Shotton J., Davison A. J., Fitzgibbon A.: Kinectfusion: Real-time dynamic 3d surface reconstruction and interaction. In *ACM SIGGRAPH 2011 Talks* (New York, NY, USA, 2011), SIGGRAPH '11, Association for Computing Machinery. URL: https://doi.org/10.1145/2037826.2037857, doi:10.1145/2037826.2037857. 1

[IZN*16] Innmann M., Zollhöfer M., Niessner M., Theobalt C., Stamminger M.: Volumedeform: Real-time volumetric non-rigid reconstruction, 2016. doi:10.48550/ARXIV.1603.08161. 1

[Low04] Lowe D. G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2 (Nov. 2004). doi:10.1023/B:VISI.0000029664.99615.94. 1

[LSS20] Lombardi M., Savardi M., Signoroni A.: Deep-learning Alignment for Handheld 3D Acquisitions: A new Densematch Dataset for an Extended Comparison. In *Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference* (2020), Biasotti S., Pintus R., Berretti S., (Eds.), The Eurographics Association. doi:10.2312/stag.20201244. 1

[LSS21a] Lombardi M., Savardi M., Signoroni A.: Cross-domain assessment of deep learning-based alignment solutions for real-time 3d reconstruction. *Computers & Graphics 99* (2021), 54–69. doi:https://doi.org/10.1016/j.cag.2021.06.011. 1, 2

[LSS21b] Lombardi M., Savardi M., Signoroni A.: Densematch: a dataset for real-time 3d reconstruction. *Data in Brief 39* (2021), 107476. doi:https://doi.org/10.1016/j.dib.2021.107476. 2

[NIH*11] Newcombe R. A., Izadi S., Hilliges O., Molyneaux D., Kim D., Davison A. J., Kohi P., Shotton J., Hodges S., Fitzgibbon A.: Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality* (2011), pp. 127–136. doi:10.1109/ISMAR.2011.6092378. 1

[NLD11] Newcombe R. A., Lovegrove S. J., Davison A. J.: Dtam: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision* (2011), pp. 2320–2327. doi:10.1109/ICCV.2011.6126513. 1

[NZIS13] Niessner M., Zollhöfer M., Izadi S., Stamminger M.: Real-time 3d reconstruction at scale using voxel hashing. *ACM Trans. Graph. 32*, 6 (Nov. 2013). doi:10.1145/2508363.2508374. 1

[PKMR15] Prisacariu V. A., Kähler O., Murray D. W., Reid I. D.: Real-time 3d tracking and reconstruction on mobile phones. *IEEE Transactions on Visualization and Computer Graphics 21*, 5 (2015), 557–570. doi:10.1109/TVCG.2014.2355207. 1

[QYSG17] Qi C. R., Yi L., Su H., Guibas L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413* (2017). 1

[RL01] Rusinkiewicz S., Levoy M.: Efficient variants of the ICP algorithm. In *Proceedings Third International Conference on 3-D Digital Imaging and Modeling* (5 2001), pp. 145–152. doi:10.1109/IM.2001.924423. 1

[VBS18] Volonghi P., Baronio G., Signoroni A.: 3d scanning and geometry processing techniques for customised hand orthotics: an experimental assessment. *Virtual and Physical Prototyping 13*, 2 (2018), 105–116. doi:10.1080/17452759.2018.1426328. 1

[WSMG*16] Whelan T., Salas-Moreno R. F., Glocker B., Davison A. J., Leutenegger S.: Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research 35*, 14 (2016), 1697–1716. URL: https://doi.org/10.1177/0278364916669237, arXiv:https://doi.org/10.1177/0278364916669237, doi:10.1177/0278364916669237. 1

[XJZ*21] Xiang X., Jiang H., Zhang G., Yu Y., Li C., Yang X., Chen D., Bao H.: Mobile3dscanner: An online 3d scanner for high-quality object reconstruction with a mobile device. *IEEE Transactions on Visualization and Computer Graphics 27*, 11 (2021), 4245–4255. doi:10.1109/TVCG.2021.3106491. 1

[ZPK18] Zhou Q.-Y., Park J., Koltun V.: Open3D: A modern library for 3D data processing. *arXiv:1801.09847* (2018). 2