Evaluating deep learning based stereo matching geometric accuracy and generalization properties*

Manfred Klopschitz, Gerald Lodron, Matthias Rüther
JOANNEUM RESEARCH Forschungsgesellschaft mbH, 8010 Graz, Austria

Abstract Convolutional neural networks define state of the art algorithms for many computer vision and image recognition tasks in recent years. The same is true for dense stereo matching, deep learning based networks outperform classical approaches on public benchmark data sets. A closer look on standard data sets used to train and evaluate these methods relveals that they consist of mostly small baseline image pairs with little perspective effects and occlusions. We investigate the geometric accuracy and generalization abilities of a state of the art stereo approach in a different setting. As real world application of stereo matching results a volumetric fusion pipeline is used. The pipeline fuses depth maps to generate a 3D model and is tracking camera poses using the model, geometric errors are accumulated over thousands of frames and result in 3D model deviations. As baseline method we use a well tested traditional semi-global stereo method that is integrated into a highly accurate and robust projected texture stereo system with efficient volumetric integration, enabling the easy capture of accurate 3D models of indoor scenes. Our stereo method is specifically optimized for random dot projection patterns, delivering complete and robust results. Our first results verify that the training data and benchmarking data based superior results of deep learning based stereo methods also generalize to a different use case in practice.

Introduction Stereo matching is a classical topic in computer vision where the goal is to compute the displacement between corresponding pixels in a pair of rectified images, referred to as disparity. The disparity image can be triangulated into metric data using the geometric calibration of the stereo system. Recently, convolutional neural network (CNN) based stereo matching methods have taken the accuracy of disparity estimation to a new level with the help of large training datasets. The emergence of reliable and high-quality consumer depth cameras has sparked significant research in the use of depth images. This has resulted in innovative approaches for efficiently integrating video-rate depth image streams into consistent 3D models. In this work we compare the metric accuracy of CNN based state of the art stereo matching to classical semi global matching (SGM) using a well-calibrated stereo system with depth image fusion and Simultaneous Localization and Mapping (SLAM) techniques commonly used with consumer-grade depth sensors. One such consumer-grade depth sensor is the Kinect camera, which utilizes structured light to provide reliable and complete depth images across a wide range of scene conditions. This has led to the development of efficient methods for fusing the depth data into a single model and using it directly for tracking in SLAM [5].

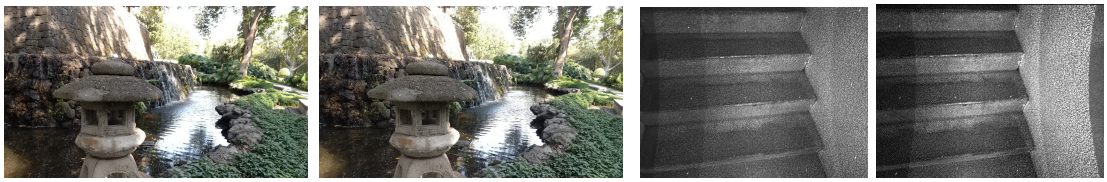
Our stereo setup employs an infrared (IR) random dot projection pattern, which allows for dense depth images that are largely independent of scene texture. We also utilize a well-calibrated stereo setup with global shutter image sensors. Our stereo matching is optimized for random dot patterns to obtain dense depth maps, which are then integrated into a consistent 3D model using recent volumetric fusion methods designed for consumer depth sensors.

As CNN based stereo algorithm [4] is used. The classical approach is represented by an implementation of SGM [1] that has been optimized to handle random dot projection patterns. The optimizations and a complete system overview are given in [3].

Stereo data The CNN based stereo algorithm has been trained on data sets like [2] and has not been optimized and retrained on our data, a typical image pair can bee seen in Figure 1, the mean pixel shift is only a few pixels, all the data sets consist of low disparity image pairs. In contrast an image pair from our system is also shown in Figure 1. The mean pixel shift is around $200px$ and contains larger perspective distortions and occlusions. Furthermore the training data does not contain random dot speckle image pairs. Adding a larger pixel shift would be of course trivial for the training data but the real difference is that larger baseline stereo images contain also more perspective change and are therefore harder to match. The only modification made to [4] was to add backmatching, i.e. also reverse the image pairs and filter for regions with inconsistent results. This filters out occluded areas.

Results The random dot projection images are designed to be well suited for stereo matching and especially well suited to work with our SGM implementation. Due to the random dot pattern the assumption that a large gray value change between two adjacent pixels likely correspond to a change of the objects depth is not valid for such imagery. The Census-based matching cost function is very well defined if a random dot (or a part of it) is located in the matching kernel. If this is not the case, all gray values are homogeneous due to the untextured scene and the Census transform generates a more or less random binary pattern. To optimize for this situation we use the so-called modified Census (MCensus)

*Acknowledgements** This work was supported by Land Steiermark within the research initiative "Digital Material Valley Styria" and FFG Contract No. 891793 "champI4.0ns".



(a) Training data [2] example image pair

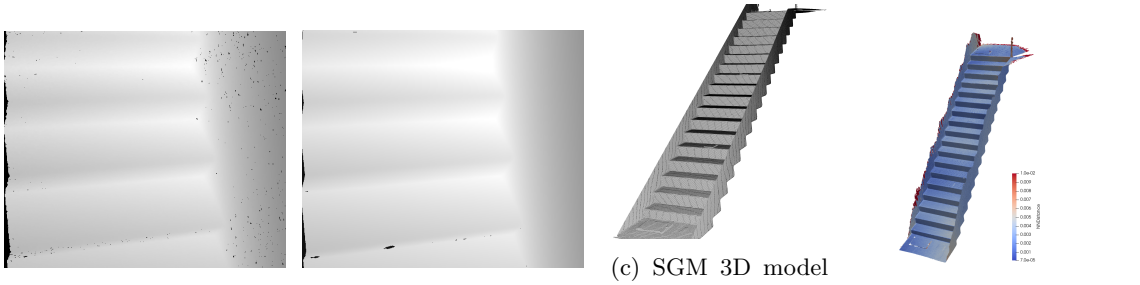
(b) IR random dot projection example image pair

Figure 1: Comparison of typical training data with very small baseline and our wide baseline IR projection image data.

transform[6] , all pixels are compared to the mean value of the current kernel instead to the central value, such that a more distinct cost value can be extracted. The SGM parameters have been optimized on this random dot image data.

In contrast the CNN based method is trained on generic image data and not optimized for this use case. Figure 2 shows a comparison of SGM and CNN depth images. The results are pretty similar with a slight advantage in overall density for the CNN based method. The difference is hardly notacible, accumulating all 1700 image frames into a 3D model sums up the differences and can make a bias or problem with the CNN based method visible.

The depths maps are fused using a simple volumetric approach [5] without bundle adjustment or other global optimization. The lack of this global optimization makes the computed camera poses only dependend on the stereo information. Given the pose of a new depth map relative to the 3D model, each frame is added to the reconstruction by computing a running average of the truncated signed distance function representation of the observed surfaces. The current sensor pose is estimated by aligning the new input depth map with the model. This alignment is implemented by iterative closest point matching of the input image and a surface prediction obtained from ray casting the truncated signed distance function representation of the model. The results are basically the same also for the 3D model as is shown with a registration of the SGM model and the CNN model differences in Figure 2.



(a) SGM stereo result

(b) CNN stereo result

(c) SGM 3D model after fusion with our optimized pipeline

(d) Difference SGM and CNN model

Figure 2: Resulting depth maps with SGM and CNN based stereo matching. The depth maps are integrated into a 3D model using a standard volumetric fusion and pose estimation mapproach. The scene in this example has a height of $2.3m$ and a diagonal length of $5.4m$. The difference between the SGM and CNN 3D model is under 2mm for most points.

Conclusion and outlook We verified that state of the art CNN based stereo matching can transfer the impressive benchmark results also to other use cases and that the generalization properties are promising. In a setting that is highly optimized for geometric accuracy no drawbacks were identified. The implications of these results are that the CNN based matchers can be further trained on these device specific random dot speckle images or that an active projection might not be necessary to get results of similar robustness. Furthermore we will investigage scenes where the CNN based matching is supposed to outperform SGM like fine structures and more pixel accurate depth edges. Fine structures and edges are also distorted by the volumetric fusion approach so this part of 3D model creation and tracking could also be optimized for more accurate depth data.

[1] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):328–341, 2008.

[2] Y. Hua, P. Kohli, P. Uplavikar, A. Ravi, S. Gunaseelan, J. Orozco, and E. Li. Holopix50k: A large-scale in-the-wild stereo image dataset. *arXiv preprint arXiv:2003.11172*, 2020.

[3] M. Klopschitz, R. Perko, G. Lodron, G. Paar, and H. Mayer. Projected texture fusion. In *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*, pages 109–114. IEEE, 2017.

[4] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022.

[5] Richard Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Andrew Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *10th IEEE International Symposium on Mixed and Augmented Reality (Proceedings of ISMAR 2011)*, Oct. 2011.

[6] B.-S. Shin, D. Caudillo, and R. Klette. Evaluation of two stereo matchers on long real-world video sequences. *Pattern Recognition*, 48(4):1113–1124, 2015.