Scene-Adaptive Disparity Estimation in Depth Sensor Networks Using Articulated Shape Models

Samuel Zeitvogel, Johannes Wetzel, and Astrid Laubenheimer

Intelligent Systems Research Group (ISRG)
Karlsruhe University of Applied Sciences
{samuel.zeitvogel,johannes.wetzel,astrid.laubenheimer}@hs-karlsruhe.de

Abstract. In this work a scene-adaptive approach for disparity estimation in depth sensor networks is presented. Our approach makes use of a priori scene knowledge to improve the low-level 3D data acquisition. We fit an articulated shape model to the given 3D data and leverage the resulting high-level scene information to make the estimation of disparities more robust to local ambiguities. We present early qualitative results to show the applicability of our method.

Keywords: scene-adaptive sensor networks, guided stereo matching, context-ware disparity estimation, articulated human shape models, stereo matching.

1 Introduction

Due to the emergence of commodity depth sensors many classical computer vision tasks are employed on networks of multiple depth sensors e.g. people detection [1] or full-body motion tracking [2]. Existing methods approach these applications using a sequential processing pipeline where the depth estimation and inference are performed on each sensor separately and the information is fused in a post-processing step. In previous work [3] we introduce a scene-adaptive optimization schema, which aims to leverage the accumulated scene context to improve perception as well as post-processing vision algorithms (see Fig. 1). In this work we present a proof-of-concept implementation of

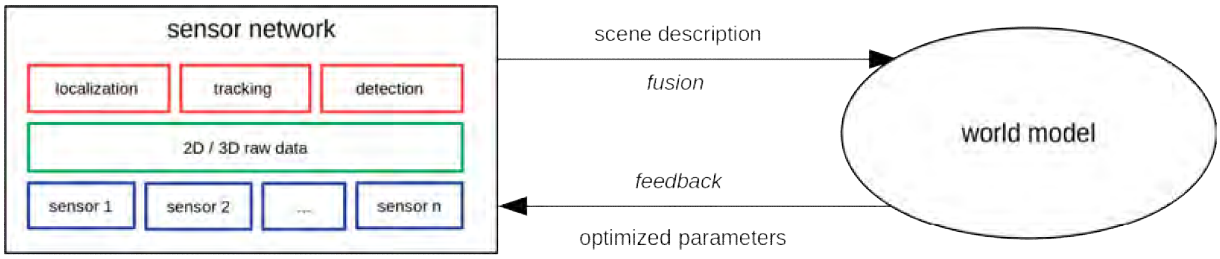

Fig. 1: Scene-Adaptive optimization introduced by [3].

the scene-adaptive optimization methods proposed in [3] for the specific task of stereo-matching in a depth sensor network. We propose to improve the 3D data acquisition step with the help of an articulated shape model, which is fitted to the acquired depth data. In particular, we use the known camera calibration and the estimated 3D shape model to resolve disparity ambiguities that arise from repeating patterns in a stereo image pair. The applicability of our approach can be shown by preliminary qualitative results.

(a) Scan (b) Shape model (c) Shoulder joints

Fig. 2: Laser scan (a), articulated shape model (b) and corresponding shoulder joints (c).

2 Related Work

In previous work [3] we introduce a general framework for scene-adaptive optimization of depth sensor networks. It is suggested to exploit inferred scene context by the sensor network to improve the perception and post-processing algorithms themselves. In this work we apply the proposed ideas in [3] to the process of stereo disparity estimation, also referred to as stereo matching.

While stereo matching has been studied for decades in the computer vision literature [4, 5] it is still a challenging problem and an active area of research. Stereo matching approaches can be categorized into two main categories, *local* and *global* methods. While local methods, such as block matching [6], obtain a disparity estimation by finding the best matching point on the corresponding scan line by comparing local image regions, global methods formulate the problem of disparity estimation as a global energy minimization problem [7]. Local methods lead to highly efficient real-time capable algorithms, however, they suffer from local disparity ambiguities. In contrast, global approaches are able to resolve local ambiguities and therefore provide high-quality disparity estimations. But they are in general very time consuming and without further simplifications not suitable for real-time applications.

The semi-global matching (SGM) introduced by Hirschmuller [8] aggregates many feasible local 1D smoothness constraints to approximate global disparity smoothness regularization. SGM and its modifications are still offering a remarkable trade-off between the quality of the disparity estimation and the run-time performance.

More recent work from Poggi et al. [9] focuses on improving the stereo matching by taking additional high-quality sources (e.g. LiDAR) into account. They propose to leverage sparse reliable depth measurements to improve dense stereo matching. The sparse reliable depth measurements act as a prior to the dense disparity estimation. The proposed approach can be used to improve more recent end-to-end deep learning architectures [10, 11], as well as classical stereo approaches like SGM.

This work is inspired by [9], however, our approach does not rely on an additional LiDAR sensor but leverages a priori scene knowledge in terms of an articulated shape model instead to improve the stereo matching process.

3 Approach

3.1 Experimental Setup

We set up four stereo depth sensors with overlapping fields of view. The sensors are extrinsically calibrated in advance, thus their pose with respect to a world coordinates system

is known. The stereo sensors are pointed at a mannequin and capture eight greyscale images (one image pair for each stereo sensor, the left image of each pair is depicted in Fig. 3a). For our experiments we use a high-quality laser scan of the mannequin as ground truth. We assume that the proposed algorithm has access to an existing shape model that can express the observed geometry of the scene in some capacity. In our experimental setup, we assume a shape model of a mannequin with two articulated shoulders and a slightly different shape in the belly area of the mannequin (see Fig. 2). In the remaining section we use the provided shape model to improve the depth data generation of the sensor network.

3.2 Semi-global Block Matching

First, we estimate the disparity values of each of the four stereo sensors with SGM without using the human shape model. Let p denote a pixel and q denote an adjacent pixel. Let d denote a disparity map and d_p,d_q denote the disparity at pixel location p and q. Let $\mathcal{P}$ denote the set of all pixels and $\mathcal{N}$ the set of all adjacent pixels. Then the SGM cost function can be defined as

$$E(d) = \sum_{p \in \mathcal{P}} D(p, d_p) + \sum_{(p,q) \in \mathcal{N}} R(p, d_p, q, d_q), \tag{1}$$

where $D(p, d_p)$ denotes the matching term (here the sum of absolute differences in a 7×7 neighborhood) which assigns a matching cost to the assignment of disparity d_p to pixel p and $R(p, d_p, q, d_q)$ penalizes disparity discontinuities between adjacent pixels p and q. In SGM the objective given in (1) is minimized with dynamic programming, leading to the resulting disparity map

$$\bar{d} = \arg\min_{d} E(d). \tag{2}$$

As input for the shape model fitting we apply SGM on all four stereo pairs leading to four disparity maps as depicted in Fig. 4a.

3.3 Shape Model Fitting

To be able to exploit the articulated shape model for stereo matching we initial need to fit the model to the 3D data obtained by classical SGM as described in 3.2. To be more robust to outliers we do only use disparity values from pixels with high contrast and transform them into 3D point clouds. Since we assume that the relative camera poses are known, it is straight forward to merge the resulting point clouds in one world coordinate system. Finally the shape model is fitted to the merged point cloud by optimizing over the shape model parameters, namely the pose of the model and the rotation of the shoulder joints. We use an articulated mannequin shape model in this work as a proxy for an articulated human shape model (e.g. [2]) as proof-of-concept and plan to transfer the proposed approach on real humans in future work.

3.4 Synthetic Disparity Maps

Once the model parameters of the shape model are obtained we can reproject the model fit to each sensor view by making use of the known projection matrices. Fig. 3b shows the rendered wireframe mesh of the fitted model as an overlay on the camera images. For our guided stereo matching approach we then need the synthetic disparity map which can

be computed from the synthetic depth maps (a byproduct of 3D rendering). We denote the synthetic disparity image by d^{synth}. One synthetic disparity image is created for each stereo sensor, see Fig. 4b.

3.5 Model-based Guided Stereo Matching

In the final step we exploit the existing shape model fit, in particular the synthetic disparity image d^{synth} of each stereo sensor and combine it with SGM (inspired by guided stereo matching [9]). Our augmented objective is defined as

$$E'(d) = \sum_{p \in \mathcal{P}} D'(p, d_p) + \sum_{(p,q) \in \mathcal{N}} R(p, d_p, q, d_q), \tag{3}$$

with

$$D'(p, d_p) = \begin{cases} D(p, d_p) & \text{if } |d_p^{\text{synth}} - d_p| \leq 3 \\ \infty & \text{else.} \end{cases} \tag{4}$$

The introduced objective is very similar to SGM and can be minimized in a similar fashion leading to the final disparity estimation in our scene-adaptive depth sensor network

$$\hat{d} = \arg\min_d E'(d). \tag{5}$$

To summarize our approach, we exploit an articulated shape model fit to enhance SGM with minor adjustments.

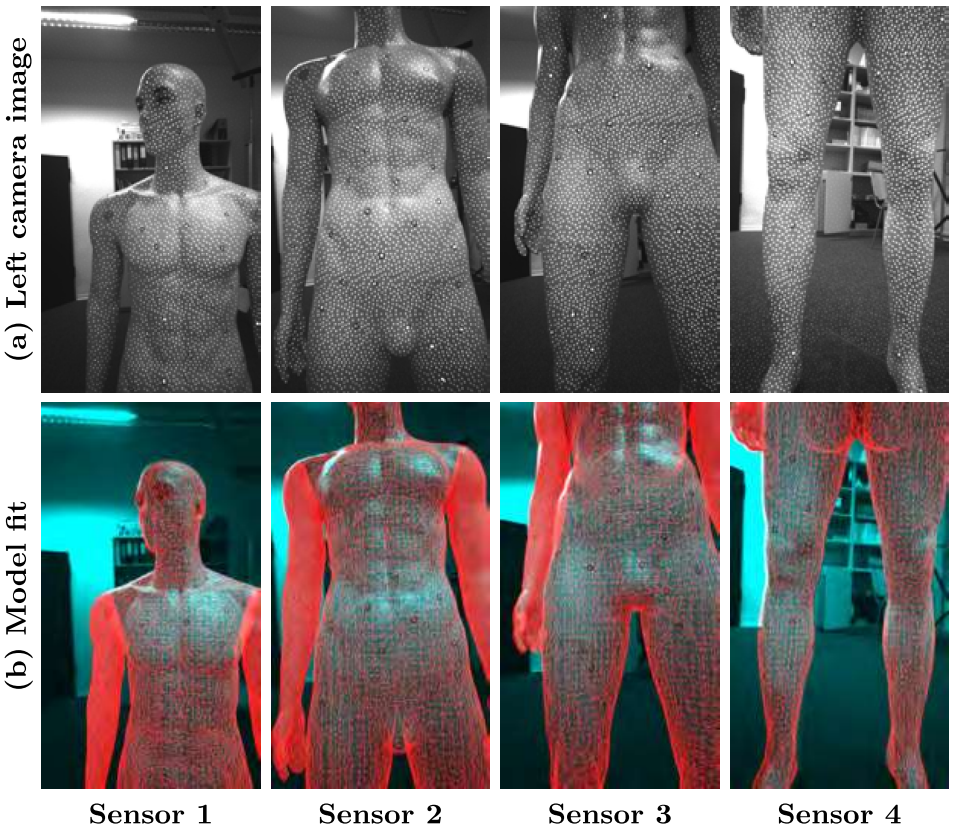

Fig. 3: Left camera images (a) and fitted shape model projected on camera images (b).

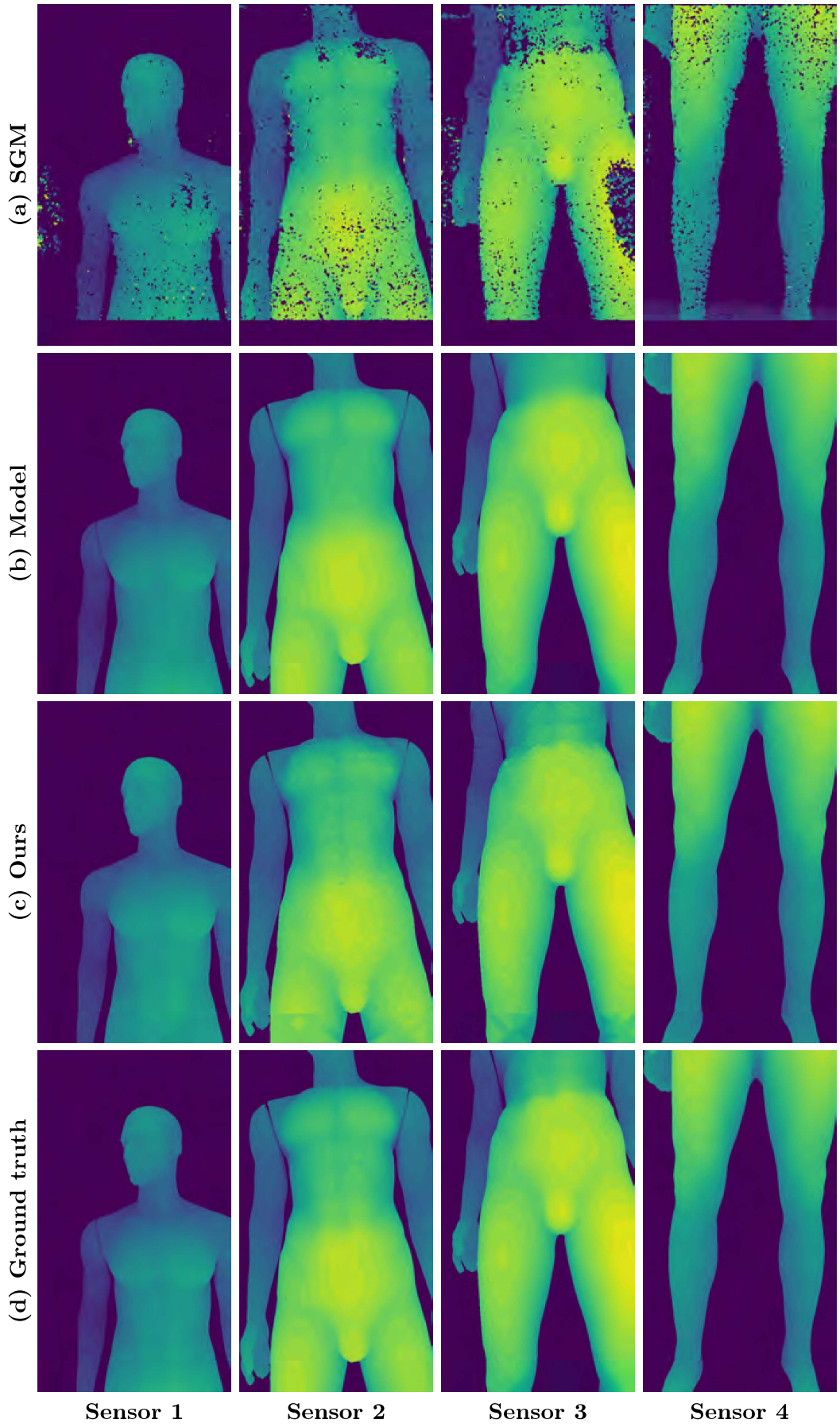

Fig. 4: Comparison of obtained disparity maps for all sensor views.

4 Qualitative Evaluation

To show the applicability of our approach we present preliminary qualitative results. The results are depicted in Fig. 4. Using SGM without exploiting the provided articulated shape model leads to reasonable results, but the disparity map is very noisy and no clean silhouette of the mannequin is extracted (see Fig. 4a). Fitting our articulated shape model to the data leads to clean synthetic disparity maps as shown in Fig. 4c, with a clean silhouette. In the belly area the synthetic model disparity map (Fig. 4b) does not agree with the ground truth (Fig. 4d). The articulated shape model is not general enough to explain the recorded scene faithfully. Using the guided stereo matching approach, we construct a much cleaner disparity map than SGM. In addition, the approach takes the current sensor data into account and exploits an existing articulated shape model.

5 Conclusion

In this work we have proposed a method for scene-adaptive disparity estimation in depth sensor networks. Our main contribution is the exploitation of a fitted human shape model to make the estimation of disparities more robust to local ambiguities. Our early results indicate that our method can lead to more robust and accurate results compared to classical SGM. Future work will focus on a quantitative evaluation as well as incorporating sophisticated statistical human shape models into our approach.

References

1. Wetzel, J., Laubenheimer, A., Heizmann, M.: Joint probabilistic people detection in overlapping depth images. IEEE Access **8** (2020) 28349–28359
2. Zeitvogel, S., Laubenheimer, A.: Towards end-to-end 3d human avatar shape reconstruction from 4d data. In: 2018 International Symposium on Electronics and Telecommunications (ISETC), IEEE (2018) 1–4
3. Wetzel, J., Zeitvogel, S., Laubenheimer, A., Heizmann, M.: Scene-adaptive optimization scheme for depth sensor networks. In: CERC. (2019) 111–116
4. Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Proc. - IEEE Work. Stereo Multi-Baseline Vision, SMBV 2001 (1) (2001) 131–140
5. Brown, M.Z., Burschka, D., Hager, G.D.: Advances in computational stereo. IEEE Trans. Pattern Anal. Mach. Intell. **25**(8) (2003) 993–1008
6. Roma, N., Santos-Victor, J., Tomé, J.: A Comparative Analysis of Cross-Correlation Matching Algorithms Using a Pyramidal Resolution Approach. (2002) 117–142
7. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. In: Proc. IEEE Int. Conf. Comput. Vis. Volume 1., IEEE (1999) 377–384
8. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2) (2008) 328–341
9. Poggi, M., Pallotti, D., Tosi, F., Mattoccia, S.: Guided stereo matching. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. **2019-June** (2019) 979–988
10. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. **2016-December** (2016) 4040–4048
11. Chang, J.R., Chen, Y.S.: Pyramid Stereo Matching Network. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2018) 5410–5418