Scene-Adaptive Optimization Scheme for Depth Sensor Networks

Johannes Wetzel[1], Samuel Zeitvogel[1], Astrid Laubenheimer[1], and Michael Heizmann[2]

[1] Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, Karlsruhe, Germany
{johannes.wetzel,samuel.zeitvogel,astrid.laubenheimer}@hs-karlsruhe.de
[2] Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
michael.heizmann@kit.edu

Abstract. In this work a scheme for scene-adaptive depth sensor network optimization is presented. We propose to fuse the knowledge inferred by the sensor network into a common world model while at the same time exploiting this knowledge to improve the perception and post processing algorithms themselves. Moreover, we show how our optimization scheme can be applied to improve the use cases of disparity estimation as well as people detection with multiple depth sensors.

Keywords: depth sensor networks · context aware · knowledge based optimization · scene-adaptive · optimization

1 Introduction

Low cost commodity depth sensors are an emerging technology and are applied to a broad field of applications such as people detection and tracking, 3D reconstruction or emergency detection in an ambient assisted living context. However, depth sensor networks as well as modern vision algorithms have many parameters and require fine-tuned, scene-specific configurations to achieve optimal performance. Due to strongly varying scenes and changing conditions at run time it is very challenging to fine-tune those parameters manually in real world applications. To overcome the problem of scene-specific manual (re)configuration of depth sensor networks, we propose a scene-adaptive scheme which exploits the scene knowledge to improve perception and post processing vision algorithms. Our objective is not only to tune the given parameters but also to improve the vision algorithms, such as stereo block matching, detection or tracking by explicit exploitation of the scene knowledge, e.g. by building scene-specific object models. Therefore, we fuse the knowledge inferred from the sensor network into a common world model, representing our current context knowledge. This knowledge is then fed back to optimize sensor parameters and algorithms to improve the performance of a sensor network at run time.

2 Related work

The configuration of video sensor networks in the context of video surveillance has been widely studied in the literature. In [13] a general overview of the different aspects of sensor network reconfiguration is given. Rinner et al. [12] focus on the aspect of configuration of smart camera networks in the context of video surveillance. They review the configuration for a specific analysis task and evaluate different configuration methods. In [8] a flexible uncertainty model is presented to reconfigure the sensor network with the objective to optimize the detection performance. Fischer et al. [4] give an overview of intelligent surveillance systems, analyzing the information flow between sensors, world model and inference algorithms. In [14] an overview to visual sensor networks is given. However, prior work focuses on monocular camera networks and employs parameter reconfiguration. In contrast, our work deals with depth sensor networks and proposes a scheme for explicit exploitation of the given scene knowledge. This includes conventional parameter reconfiguration methods as well as methods that construct and use sophisticated world models to improve the integrated algorithms of sensor networks at run time.

3 Scene-adaptive sensor network optimization

In this section we present a scheme for scene-adaptive sensor network optimization. The general information flow in a depth sensor network is depicted in Fig. 1 and separated into five different abstraction layers. The **sensing** layer

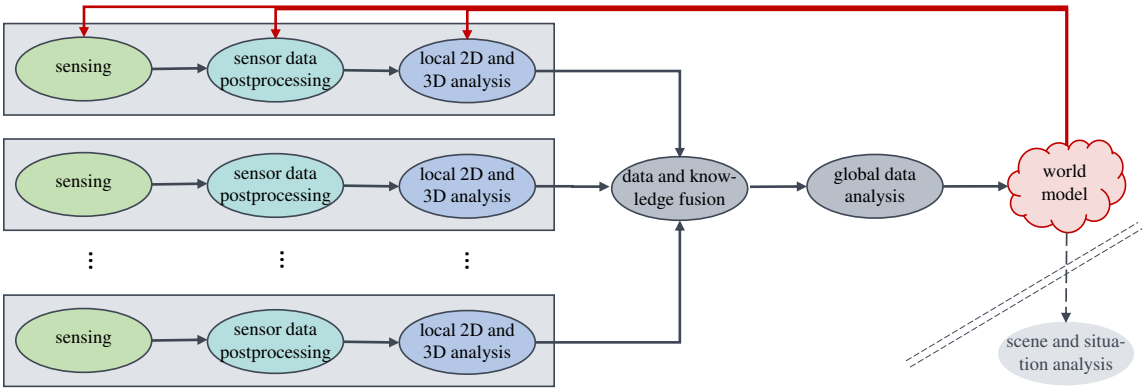

Fig. 1. Information flow in a depth sensor network with scene-adapative optimization strategy.

contains low-level methods related to the raw sensor measurement such as synchronization, calibration and image acquisition. In the **sensor data post processing** layer depth estimation algorithms (e.g. stereo block matching), filtering and low-level feature extractors are included. **Local data analysis** covers high level vision algorithms which take the RGB-D data as input such as segmentation, recognition, object detection, local 3D object and scene reconstruction or

tracking of objects. Based on the results of the **data and knowledge fusion**, the **global data analysis** layer includes methods which make use of the fused information of multiple sensors of the network. Examples are 3D scene reconstruction, 3D object localization and global tracking. The sensor network infers information about the scene across abstraction levels. Over time, information is fused into a common world model which represents the current scene knowledge. While a world model can be used to do e.g. scene and situation analysis, we use it to optimize the parameters of each individual sensor online and support the data analysis methods e.g. by building scene-specific object models gradually.

3.1 Knowledge representation

The employed knowledge representation within the world model has to be expressive to solve the high-level task of the sensor network and the optimization of the sensor network itself. The fusion layer might provide sensor data as well as locally derived high-level knowledge and the world model therefore might need to cover low-level data up to high level information. Taking these aspects into account, several existing approaches for knowledge representations are qualified to serve as world model. For most tasks and networks, a world model consisting of geometric and semantic scene descriptions will be suitable. Geometric scene knowledge thereby encompasses information about the objects contained in the scene and their properties. This includes the object class (e.g. humans, furnature, floor plan), the object location and orientation in a global world coordinate system, dynamic properties e.g. a motion model, shape, material. Examples for such a world model are object oriented world models [2, 5]. In order to enhance the quality of the world model, a knowledge base consisting of preprocessed information or prior knowledge can be used. This includes morphable shape models [3] for different object classes as well as common recognition, detection and segmentation models [18] which are applied on image and 3D data, e.g. RGB-D data, point clouds, voxels or triangulated surfaces [1]. In terms of semantic knowledge Fuzzy Metric Temporal Logic and Situation Graph Trees [11] or ontologies [10] can be incorporated. The semantic description might be data driven, e.g. Hartz and Neumann [6] use a scene interpretation system [7] and learn ontological concept descriptions from data.

3.2 Optimization possibilities

Depth sensor networks involve multiple algorithms which leads to a large amount of parameters. In this section we give an overview of parameters and methods which are suitable for automatic scene-adaptive sensor optimization. We assume that a suitable knowledge base (see section 3.1) exists and focus on algorithm and parameter optimization. Following our layered scheme, we categorize the optimization targets into three major categories, see Fig. 2. **Sensing** parameters have a direct impact on the measurement quality. Parts of this category have already been addressed. Auto exposure is state-of-the-art for decades in consumer cameras, but sophisticated scene models [17] can improve the result

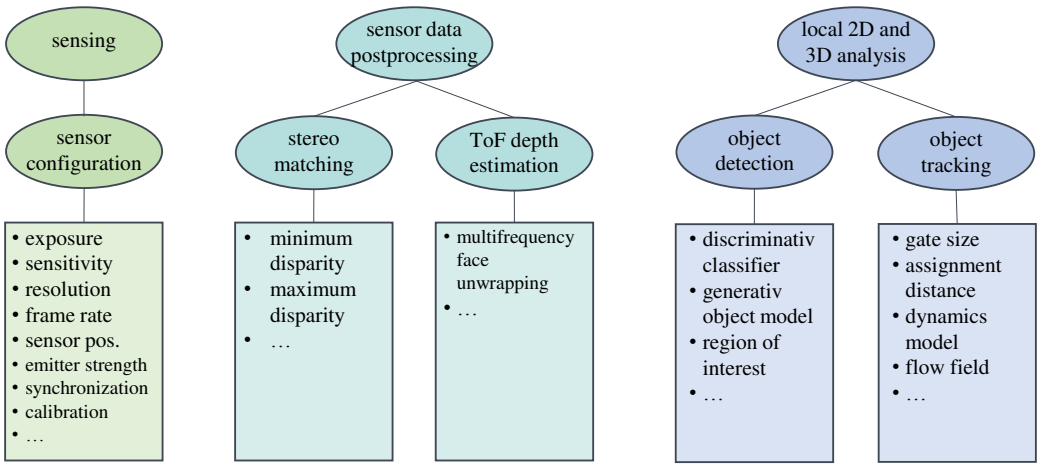

Fig. 2. Non-exhaustive taxonomy of building blocks within a depth sensor network which are suitable for scene-specific optimization.

e.g. by taking only the pixel intensities near regions of interest into account. **Sensor data post processing** methods vary highly between different depth sensing technologies. The depth estimation of a stereo sensor can be improved by setting the minimum and maximum observable disparity based on geometric scene knowledge. In section 4.1 a approach for the task of scene-adaptive disparity estimation is presented with an exemplary knowledge representation. Many scene-adapative **local data analysis** methods have already been published. Yang et al. [16] learn global appearance and motion models to improve multiple target tracking. Masksai et al. [9] propose a context-aware optimization strategy for multi object tracking. They learn the most likely trajectory patterns with respect to a given scene layout to reduce incorrect assignments between detections and tracks. In 4.2 we show how the task of people detection can be optimized in a scene-specific fashion.

4 Application

In this section we show the applicability of our scheme on two exemplary use cases.

4.1 3D model based disparity estimation

Our knowledge representation contains sensor knowledge in the form of a camera model and existing camera calibration parameters π, scene geometry using a ground plane assumption $P(h) \subset \mathbb{R}^3$ and a 3D morphable human surface model parameterized by β. Scene semantics are represented as segmentations of a single human s_h and the ground plane s_g in the image. Let $D_\pi(u)$ be a depth image computed using the estimated disparity values u from the image pair (I_1, I_2). Classical stereo algorithms estimate the disparity values u minimizing a cost function

$$E(u) = E_{\text{photometric}}(u; I_1, I_2) + E_{\text{reg}}(u) \ , \tag{1}$$

where $E_{\text{photometric}}$ is the photometric error penalizing intensity deviation in the local neighborhood given u and E_{reg} regularizes the problem penalizing unlikely disparity values based on simple scene assumptions. We propose to employ a scene-adaptive optimization scheme reformulating (1) with

$$E_{\text{adaptive}}(u) = E_{\text{photometric}}(u; I_1, I_2) + E_{\text{model}}(u[s_h], u[s_g]; \beta, h) \ , \qquad (2)$$

where E_{model} uses our provided scene representation to measure the deviation from the estimated depth at the segmented pixel locations $u[s_h]$ and $u[s_g]$ to the explicit geometric scene representation consisting of the ground plane at height h and the human shape model parameterized by β. Scene-adaptive disparity estimation is then performed by estimating $\hat{u} = \arg\min_u E_{\text{adaptive}}(u)$. Eq.(2) can be extended in various ways, which proves the generality of the proposed approach by e.g. introducing a human motion model to enforce temporal consistency constraints.

4.2 People detection with multiple depth sensors

The sensors have a top view on the scene and a significant overlap to each other. Additionally, we assume that the sensors are intrinsically and extrinsically calibrated in advance and that the common ground plane is known. We model the presence of a person on the ground floor as a discrete grid of Bernoulli random variables $\boldsymbol{X} = (x_1, .., x_n)$, $x_i \in \{0, 1\}$ where each x_i maps to one specific ground plane grid location $\boldsymbol{g}_i \in \mathbb{R}^2$. Our goal is to infer the likelihood of a scene configuration $\boldsymbol{X}$ given current depth observations $\boldsymbol{O} = (O_1, \ldots, O_C)$ from C depth sensors. Applying Bayes' theorem and assuming that the prior factorizes as $p(\boldsymbol{X}) = \prod_{i=1}^{n} p(x_i)$ we get the posterior distribution

$$p(\boldsymbol{X}|\boldsymbol{O}) \propto p(\boldsymbol{O}|\boldsymbol{X}) \prod_{i=1}^{n} p(x_i). \qquad (3)$$

For this application we assume that the likelihood $p(\boldsymbol{O}|\boldsymbol{X})$ is given (see [15] for details on the construction of the likelihood) and only focus on the scene-adaptive choice of the prior $p(\boldsymbol{X})$. We start with an uninformative prior to make the detection of people at every location equally likely. In many real world scenes this is a crude assumption due to obstacles or preferred walking tracks which can be present in the scene. Thus, we propose to accumulate the detections over time to get the relative frequencies $\boldsymbol{H} = (h_1, \ldots, h_n)$ of the presence of people for every ground plane grid location $\boldsymbol{g}_i$ and fuse those information into the world model. This scene-specific knowledge can be used in the feedback step to continuously update the prior beliefs $p(x_i)$ accordingly to $\boldsymbol{H}$ on regular time intervals.

5 Conclusion

In the present work we have proposed a scheme for scene-adaptive optimization of depth sensor networks. We have given an analysis of relevant knowledge representations and categorized identified optimization targets. Moreover, we have

exemplarily applied our scheme on the use cases of disparity estimation as well as people detection with multiple depth sensors. Future work will include the investigation of more use cases as well as proof of concept implementations.

References

1. Ahmed, E., Saint, A., Shabayek, A.E.R., Cherenkova, K., Das, R., Gusev, G., Aouada, D., Ottersten, B.: Deep learning advances on different 3d data representations: A survey. arXiv preprint arXiv:1808.01462 (2018)
2. Bauer, A., Emter, T., Vagts, H., Beyerer, J.: Object-oriented world model for surveillance systems. In: Future Security: 4th Security Research Conference. pp. 339–345. Fraunhofer Verlag (2009)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on computer graphics and interactive techniques. pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999)
4. Fischer, Y., Beyerer, J.: A top-down-view on intelligent surveillance systems. Proc. of the 7th International Conference on Systems (c), 43–48 (2012)
5. GheŢa, I., Heizmann, M., Belkin, A., Beyerer, J.: World modeling for autonomous systems. In: Dillmann, R., Beyerer, J., Hanebeck, U.D., Schultz, T. (eds.) KI 2010: Advances in Artificial Intelligence. pp. 176–183. Springer Berlin Heidelberg (2010)
6. Hartz, J., Neumann, B.: Learning a knowledge base of ontological concepts for high-level scene interpretation. In: ICMLA. pp. 436–443. IEEE (2007)
7. Hotz, L., Neumann, B.: Scene interpretation as a configuration task (2005)
8. Kyrkou, C., Christoforou, E., Timotheou, S., Theocharides, T., Panayiotou, C., Polycarpou, M.: Optimizing the detection performance of smart camera networks through a probabilistic image-based model. IEEE Transactions on Circuits and Systems for Video Technology **8215**(c) (2017)
9. Maksai, A., Wang, X., Fleuret, F., Fua, P.: Non-markovian globally consistent multi-object tracking. In: IEEE ICCV. pp. 2544–2554 (2017)
10. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification. arXiv preprint arXiv:1612.04844 (2016)
11. Münch, D., IJsselmuiden, J., Arens, M., Stiefelhagen, R.: High-level situation recognition using fuzzy metric temporal logic, case studies in surveillance and smart environments. In: ICCV Workshops. pp. 882–889. IEEE (2011)
12. Rinner, B., Dieber, B., Esterle, L., Lewis, P.R., Yao, X.: Resource-aware configuration in smart camera networks. IEEE CVPR (1), 58–65 (2012)
13. Sanmiguel, J.C., Micheloni, C., Shoop, K., Foresti, G.L., Cavallaro, A.: Self-reconfigurable smart camera networks. IEEE Computer **47**(5), 67–73 (2014)
14. Soro, S., Heinzelman, W.: A survey of visual sensor networks. Advances in Multimedia (2009)
15. Wetzel, J., Zeitvogel, S., Laubenheimer, A., Heizmann, M.: Towards global people detection and tracking using multiple depth sensors. IEEE ISETC, Timisoara (2018)
16. Yang, B., Nevatia, R.: An online learned crf model for multi-target tracking. IEEE CVPR pp. 2034–2041 (2012)
17. Yang, H., Wang, B., Vesdapunt, N., Guo, M., Kang, S.B.: Personalized attention-aware exposure control using reinforcement learning **14**(8), 1–17 (2018)
18. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. arXiv preprint arXiv:1807.05511 (2018)