Ground Truth Generation

Daniel Kondermann, Heidelberg Collaboratory for Image Processing

Problem Specification and Relevance

Low-level vision severely lacks meaningful performance evluations. For example, optical flow estimation algorithms do not yet deliver sufficiently good results for e.g. driver assistance or medical imaging. To date, more than 1500 papers deal specifically with improving these methods. While the number of publications grows exponentially, almost none of the proposed methods could yet be evaluated. Hence, engineers are not able to select the one approach which best works for their application. A major reason for this problem is a lack of carefully measured reference data, also called ground truth.
To ultimately answer questions about the suitability of an image processing algorithm for a given application, I pose three main questions: How can we (1) *cost-effectively* create (2) *large* amounts of (3) *accurate* ground truth?
This entails a number of more detailed and largely unanswered questions, such as: Can we trust synthetically rendered depth or RGB sequences? How do we obtain geometry, materials, textures and animations for such datasets? Do we have good enough camera models (ToF, stereo, RGB, etc) to synthesize realistic noise? Can we trust human annotations as ground truth? Can we use measurement sciences to create ground truth for real scenes? In general, for which applications which accuracy of ground truth do we need? How can applications deal with missing and inaccurate ground truth data? Can we bootstrap ground truth with vision methods using more data? What constitutes a good ground truth dataset? When do we have enough ground truth? Given a real application, which ground truth dataset is the best for studying the performance? Can we enable anybody to quickly generate ground truth for her own application? Which ground truth do we need in addition to the existing datasets?

Application Areas and Implications

Low-level vision such as stereo, optical flow and feature-tracking is the basis for all vision-based applications. Ground truth generation for low-level vision is a challenging task but mandatory for any advances in the field. Applications are security-relevant fields such as driver assistance systems, medical imaging, robotics, surveillance and more. Under ideal conditions an engineer of a vision system should be able to read a specification sheet with a number of key performance measures indicating under which constraints a method can be applied to which kind of data. *Appropriate* ground truth would enable meaningful performance evaluations which in turn would give guarantees with respect to applicability, quality of results, graceful degradation, time, energy consumption and more.

Speculative Cause of the Problem

An unambiguous cause of the problem of ground truth generation cannot easily be found. In a recent book draft [1], Burfoot picks up on the points of [3]. According to the author, *"The weakness of evaluation in computer vision is strongly related to the fact that the field does not conceive of itself as an empirical science. [...] Instead [...], vision researchers see themselves as producing a suite of tools."* (p.103).
He also sees similarities to historical problems in other fields of science such as physics and chemistry: *"It is almost as if, by viewing birds, researchers of an earlier age anticipated the arrival of artificial flight, and proposed to pave the way to that application by developing artificial feathers."*(p. 106) *"The argument of this book, then, is that the conceptual obstacle hindering progress in computer vision is simply a reincarnation of one that so long delayed the development of physics and chemistry."* (p. 108) *'"The difference is that physicists can eventually determine which explanation is the best. One crucial aspect of the success of the field of physics is that physicists are able to build on top of their predecessors' work."* (p. 105)

I think that much truth lieas in the fact that computer vision is a comparably young field of research which historically originates from artifical intelligence and (notably) not from measurement sciences[1]. I would guess that the past 30 years can be understood as a brainstorming phase in which any method which works more or less well to solve a more or less well-defined problem was an important scientific result and therefore worth a publication.
Ground truth generation itself is clearly an engineering task which costs a lot of time and effort with relatively little scientific joy for its creators. As a result, most researchers gladly accept any type of ground truth without asking too many of the above-mentioned questions. On the other hand, the selection of relevant data worth to be annotated with ground truth clearly is a scientific task. It is time to establish a generally accepted culture of consolidation of results, even if this means that some historically well-received papers need to be re-evaluated.

State of the Art

My background mainly lies in optical flow and stereo vision. As generation of ground truth for these methods is one of the most demanding tasks, I focus on related work in this field.
A comprehensive overview of the problem of performance evaluation (including ground truth) for optical flow is given in [5]. In 2012, three new optical flow reference datasets have been published, two of them containing ground truth [6, 2, 4]. Unfortunately, none of them contains ground truth for real-world, large-scale outdoor scenes with dynamically and independently moving objects. The reason is that no measurement devices exists to record such data with sufficiently high accuracy.
To the best of my knowledge, none of these (or any previous) publications closely investigate either the relevance, cost or the accuracy of the created datasets. This is a bold statement which I would like to defend in the context of the workshop. Furthermore, workshop discussions should also other fields of research such as stereo, matting or segmentation.

Necessary Steps

I believe that the necessary steps are to rigorously answer all the questions posed above.
This requires a general acceptance of the currently desolate state of affairs and the willingness to widely support actually reproducible research.
We also need to address the interdisciplinarity of the task: next to dissertation projects we need long-term technical positions along with close collaborations with industry partners and other scientific disciplines such as computer graphics, measurement sciences and systems engineering.
As the described questions are so numerous and diverse I do not believe that a few labs can effectively and efficiently address the problem. I therefore argue for large long-term scientific projects such as BMBF Competence Centers, DFG Forschergruppen or even Sonderforschungsbereiche.

References

[1] D. Burfoot. Notes on a new philosophy of empirical science. *Arxiv preprint arXiv:1104.5466*, 2011.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012.

[3] W. Förstner. 10 pros and cons against performance characterization of vision algorithms. In *Proc. of ECCV Workshop on Performance Characteristics of Vision Algorithms*, pages 13–29, 1996.

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR)*, Providence, USA, June 2012.

[5] D. Kondermann, S. Abraham, G. Brostow, W. Förstner, S. Gehrig, A. Imiya, B. Jähne, F. Klose, M. Magnor, H. Mayer, et al. On performance analysis of optical flow algorithms. *Outdoor and Large-Scale Real-World Scene Analysis*, pages 329–355, 2012.

[6] S. Meister, B. Jähne, and D. Kondermann. Outdoor stereo camera system for the generation of real-world benchmark data sets. *Optical Engineering*, 51, 2012.

[1] The latter field pursues a strict code of conduct when it comes to performance evaluations