Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-196-AC3, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "GRAINet: Mapping grain size distributions in river beds from UAV images with convolutional neural networks" *by* Nico Lang et al.

Nico Lang et al.

nico.lang@geod.baug.ethz.ch

Received and published: 26 August 2020

We thank the reviewer for his insightful and positive feedback. We are happy that our work is seen as showing a new exciting perspective, and as a significant and novel contribution. Thank you for the thoughtful inputs and the hints to the useful related work that will strengthen the motivation for our work. We address the major points below and will consider them for the revised version of the paper.

"I would avoid the reference to Fehr (1987)"

The main purpose of the paper is to propose and evaluate a new method to estimate

grain size distribution from raw UAV images, using convolutional neural networks. We see that the standard methodology to measure grain sizes in the field may vary between countries. Fortunately, there is no tight coupling, the CNN is agnostic and will learn to replicate the outcome of any consistent procedure for creating ground truth grain size distribution samples. The annotation strategy can easily be exchanged and the network retrained to estimate grain size distribution according to a different National or regional standard (e.g., grid sampling). However, since the line sampling by Fehr continues to be the standard field method in the German speaking world, we found our digital line sampling to be most efficient, while providing representative reference data with respect to the "gold standard" of line sampling in the field.

Introduction

Thank you for the inputs. We will incorporate the related work and will revise the introduction to clarify "why grain size data are crucial". We already describe the importance of grain size in line 29 and following.

"L 15-28. This part is not very useful."

In our opinion it is important to give a broad motivation for our research and explain its relevance for society. If the reviewer (respectively, editor) insists we will of course remove the paragraph, but we would much prefer to keep it.

"I would say that automatic grain size is much less time consuming but it is also, commonly, less accurate."

We agree and mention that the automatic grain size estimation from images is still limited in terms of accuracy when it comes to large scale applications (line 56 in the paper).

Ground truth

"Is this really a ground truth? These measurements of grains are obtained from images not from direct measurements. I understand that this can be the way for training the model, but I would not say that these are ground truth."

In machine learning terminology, the term "ground truth" refers to the data that is used to train and evaluate the model (being the upper bound the model can ever reach if it manages to perfectly replicate the annotations). To clarify the terminology, we explicitly declare in line 343, that the term "ground truth" refers to the digital line samples. In the revision we will clarify this earlier in the paper in section 3.3, to avoid any possible confusion.

Comparison with field measurements

"How field measurements were carried out should be explained in detail (in the Method section)."

We will explain that field samples were measured according to line sampling proposed by Fehr (1987). But we feel it would be too much to add another section in the methodology, given that this is not the focus of the paper, and a widely used standard practice.

"At least for those field measurements of known location, it would be crucial to show the real difference with digital line samples."

As mentioned in line 336, the location of the field measurements is not known exactly, because the field measurements were originally recorded for other purposes, within other projects. The best we can do with the available data is to evaluate the bar-level agreement between (independent) field work and our digital line sampling approach. For the training and evaluation of the CNN we consider the digital line samples to be representative (line 344). We do agree that a comparison between digital line samples and geolocalized field samples would be interesting, but since this data is not available, we cannot provide it in this paper.

"A better comparison with digital line samples should be carried out: I do not agree that ". . .overall, no bias exists between the field measurements and the digital line samples" (L 343-344; figure 6)."

To better assess the agreement between the Dm derived from field and digital line samples at the bar level, we compute the overall bias across the 22 bars with available field data (see Figure 6). The mean error (bias) amounts to -0.3 cm, which means that the digital Dm is on average slightly lower than the Dm derived from field samples. The mean absolute error is 0.9 cm. The reviewer will no doubt agree that field samples are unavoidably affected by the selected location (line 341) and also by operator bias (Wohl 1996).
Hence, we still conclude that within reasonable expectations the digital line samples are in good agreement with field samples and constitute representative training data.

Robust estimation

Figure 11 shows the estimated grain size distribution for entire gravel bars in comparison to the ground truth data. Given appropriate training data, the model estimates the grain size distributions very accurately. Due to the inherent regularisation, deep learning models tend to be fairly robust.

Manual component of the presented approach

Semi-automatic image labelling might be an alternative way to speed up the annotation process, but one would have to carefully avoid systematic algorithmic biases in the semi-automatic procedure, otherwise the CNN will almost certainly learn to faithfully reproduce those biases. Similarly, systematic behaviours of specific annotators may also be learned by the model. Ideally, training data should thus be generated by different annotators with comparable (preferably high) skill. Independent of possible biases, the automatic approach handles all samples consistently and allows for unbiased monitoring over long times, as there is no variation due to changed operators (Wohl et al.,

1996).

"Comparison with human performance (section 5.4.4). Errors are not so small, see figure 15"

The vertical axis label in Figure 15 seems to have been mangled during pdf generation. To clarify, the Y-axis in Figure 15 corresponds to the Dm, not the error. We compare the mean diameter (Dm) estimated by the CNN to the Dm from multiple annotations by different operators. This comparison requires a lot of manual labour (repeated, independent annotation by different people), thus only a small number of samples could be processed, which is not ideal for statistical analysis. However, it still gives a feeling for the human variation in the annotation process, with an average standard deviation across different operators of 0.5 cm for Dm. The max standard deviation from repeated annotation is 2.0 cm.
We correct a small mistake in line 454: The standard deviation of the labels (0.5 cm) should be compared with the root mean squared error (RMSE) of the model predictions, not with the mean absolute error MAE. Hence, regressing Dm with GRAINet yields RMSE=1.7 cm, from which 29% can be explained by the label noise in the test data."

Presence of fine material

We have already included two examples in Figure 8 b) and c) that show that the network can make robust predictions even in the case of slight disturbances caused by colmation through fine material, given such samples are provided during training.

References

Wohl, E. E., Anthony, D. J., Madsen, S. W., and Thompson, D. M. (1996). A comparison of surface sampling methods for coarse fluvial sediments. Water Resources Research, 32(10), 3219-3226.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-196, 2020.