Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-196-AC2, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Hydrology and
Earth System
Sciences
Open Access
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

Dear Patrice Carbonneau,

Thank you for your thoughtful feedback. It is motivating that you find our work innovative and the results extremely impressive and promising. Thank you for pointing out technical details, which we will incorporate in the revised version. We will refer to the additional related work, where appropriate, in the revised paper. We address your discussion points in the sections below.

Drone Flights and Image resolution

"The method would have much more impact if it could operate on the existing flight patterns (50+ meters flight altitudes) and show that you don't need the near-ground flights to get grain size distribution."

The choice of image resolution was driven by the requirement that a human annotator should be able to reliably label individual grains down to 1 cm in size. We found that a ground sampling distance of 0.25 cm is the limit for this task. To achieve that resolution, the specific UAV used in this study flew 10 m above the ground. However, the flying height depends only on the image magnification, flying at higher altitude is perfectly possible with a suitably chosen camera payload (sensor and lens). Hence, we do not see the input resolution of 0.25 cm as a serious limitation of our proposed approach, especially given the fast progress of continuously improving and ever cheaper hardware.

"The authors did try to look at the effect of resolution. But in my experience with image texture work, I have repeatedly found that downsampling an image digitally is not the same thing as acquiring the scene from a higher altitude."

We study the effect of lower input resolutions on the performance of the grain size estimation by downsampling the high-res images (with appropriate smoothing to avoid aliasing effects). To our knowledge this is the best practice and widely used e.g. in the image analysis and computer vision community to study super-resolution tasks. Our study shows that the performance of the CNN is not affected up to 2 cm ground sampling distance (downsampling factor 8), which would be the comparable resolution achieved at 80 m flying altitude with the particular low-cost UAV used in this work. We are confident that, although our synthetic downsampling may indeed miss subtle optical effects, actual images taken with 1 cm GSD would perform at least as well as simulated ones with 2 cm. Since we do not have high-altitude orthophotos for the investigated gravel bars, this is the best study we can provide.

"The authors need to choose their wording more carefully. The innovation here is that

for each patch, GRAINet gives a distribution."

We agree and will be more precise in describing the novelty with respect to your work. To summarize, our presented approach has the following contributions:

- End-to-end estimation of the grain size distribution at particular locations over an area of 1.25 m x 0.5 m
- Robust grain size distribution for entire gravel bars.
- Overall performance is invariant to the image resolution up to 2 cm ground sampling distance.
- Generic approach that also allows to map particular grain size metrics like the mean diameter with the same model architecture.
- Mapping of mean diameters <1.5 cm

Training and Overfitting

"As it stands, there is even a footnote saying it is overfit."

This seems to be a misunderstanding. To clarify, the footnote number 2 on page 4, referring to line 94, is talking about the related work by Buscombe (2019), where it appears that model selection and performance evaluation were done with the same data, which typically leads to an overfit with overly optimistic results. "Model selection" in this context means choosing at which iteration of the learning sequence the network parameters are read out. That hyper-parameter must be tuned on a validation set, without looking at the performance on the actual test data (see also Goodfellow et al. (2016), Chapter 5, section 5.3). Besides sanity-checking the training, a main purpose of monitoring the validation loss is to detect when the relatively best parameters have been found, as overfitting to the training data begins.

Contrary to our understanding of Buscombe's work, we do use the correct procedure (line 323 in the paper), i.e., our model is not overfitted and the reported performance is an unbiased estimator of the expected performance on unseen images.

"...the authors must establish that their network is well trained."

We carried out a range of careful experiments to evaluate the performance of the network, and present its performance change w.r.t. random training-test splits and w.r.t. particular training-test gravel bars. As mentioned in line 471, the latter is coupled with particular training-test imaging conditions, since each gravel bar was recorded individually, with specific imaging conditions. In our opinion this is a rather comprehensive and transparent form of evaluation. We also tried to explicitly discuss both advantages and weaknesses. If you believe there is a better way to assess model training (with the available data), perhaps you could suggest a concrete experiment?

Validation

"But what they call 'geographical cross-validation' is nothing else but bar-scale boot-strapping (sometimes called jack-knifing). It is not appropriate or reliable in this context. The main issue is that even if the authors hold out a whole bar, there remain samples from the same river. Given that river properties vary relatively slowly with geology and sometimes tributary inputs, the boot-strapped training samples will still have data that is very similar to the label data. So this is not a good test of generalisation. A much better approach to this would be to hold out an entire river."

We respectfully disagree, but we agree that this is a good discussion point and needs to be clarified in the paper.
In some geographical conditions the river properties may vary slowly, but the characteristics of the investigated bars can hardly be grouped by the river name. The relevant statistics of the investigated gravel bars are presented in Table B1 in the paper. The characteristics of the investigated gravel bars from the same river are diverse, both

quantitatively (mean Dm, Dm range in Table B1) and qualitatively (see the attached Figure 1, where tiles are grouped by river name). Not only is this due to the distance between the bars, but also due to the changing slope and the varying river bed widths in mountain environments (Reuss, Aare, Emme). Furthermore, the bars are geographically separated through tributaries (Aare, Rhone), leading to a drastic increase in the catchment areas between the bars. E.g., at the river Rhone the catchment area is more than doubled from 982 km^2 (at km 083.3) to 2485 km^2 (at km 114.0). On the other hand the sediment transport is affected by dams at the rivers Aare, Rhone, and Reuss. Finally, the characteristics may also be artificially altered, as it is nowadays common in Central Europe to replenish gravels of 2-3 cm to create spawning grounds for fish. For instance, the grain size distribution at the bar Reuss km 022.1 (and probably km 012.0) is very likely affected by such a targeted replenishment of sediment.

Hence, we are convinced that the presented hold-one-bar-out experiment is a valid setup to evaluate the generalization of the approach to new locations. In this setup, the data is exploited best, allowing the CNN to learn features invariant to the imaging conditions by providing 24 different orthophotos in each experiment. If we would hold out, e.g., the whole river Aare, not only the number of training samples would be substantially reduced, but also the diversity of imaging conditions.

In fact, within our experimental setup we already present one hold-one-river-out experiment for the river Kl. Emme, from which only one bar is included in our dataset. Even though this bar contains the largest number of digital line samples, its estimated grain size distribution fits rather well in the geographical cross-validation experiment (see Figure 17). The observed performance drop between the random and geographical cross-validation experiment for individual bars in Figure 18 is rather explained by coarse gravel bars with a large mean Dm and a wide Dm range. Seeing bars from the same river during training and testing does not seem to have an effect (for instance, Aare).

It is also important to keep in mind that data-driven approaches, like the one proposed, will only give reasonable estimates if the test data approximately matches the training

data distribution. It will not perform well for out-of-distribution (OOD) samples. Detecting such OOD samples is an unsolved problem and an active research direction.

“As it stands, I think the results show that the network is not well trained, has too many parameters or too little training data.”

We do not understand this statement. Since no further explanation is given, we cannot properly address it. We can only hypothesize that this claim is based on the misunderstanding of footnote 2, which we have discussed above.

In fact, our experimental setup quite clearly reveals in which specific cases the limited and unbalanced training data causes performance issues. Namely, our results degrade for coarse gravel bars, which is linked to the fact that only 14% of the training data have Dm > 10 cm (see lines 515-522 in the paper).

Network Architecture Description

“Why this network with residual blocks?”

After their introduction by He et al. (2016), CNNs with residual blocks quickly became the preferred design for deep image interpretation models (including many successors that no longer have the “residual” in their names). Residual connections improve gradient flow during training, because they provide shortcuts that mitigate the vanishing gradient problem. We refer to He et al. (2016), Veit et al. (2016) for more details. The model depth and capacity is empirically calibrated on the validation data for best performance on the grain size estimation task. In line with recent literature, our view is that a well-chosen loss function that matches the application objective is key, and more important than the specific choice of CNN architecture. Well-proven, contemporary architectures are rather interchangeable, and the right depth (capacity) depends on the specific task and dataset size. If a larger training set were available, an even deeper network might well perform better.

“At the very least the work needs a figure to detail the network architecture.”

We will add the illustration of the network in the revised paper (see attached Figure 2).

"It would also be helpful if the authors could show some of the activation maps in order to confirm that the network is 'seeing' reasonable elements that can conceivably lead to a regression to grain size distributions."

Thank you for the suggestion, we will add examples of activation maps to the revised paper (see attached Figure 3). We show the activation maps after the last convolutional layer, before global average pooling. Hence, each of the 21 maps corresponds to a specific bin of grain sizes, with bin 0 for the smallest grains and bin 20 for the largest ones. Light colours are low activations, darker red denotes higher activations. To harmonize the activations to a common scale [0, 1] for visualisation, we pass the maps through a softmax over bins. The activation maps intuitively make sense, with smaller grains activating the corresponding, lower bin numbers.

Comparison to past work

"They must show that this is not just a new way of doing existing jobs, it represents real progress. And to do this they must cite errors from other work and clearly demonstrate that their results are better. This should include SedNet which is the rival CNN approach. And given that the title has the term UAV, they must be much more explicit in discussing their method in relation to other UAV (and airborne) particle sizing methods. Ideally, they should try to adapt the method so that it represents real progress with respect to other UAV methods in terms of time spent in acquisition, pre-processing and final data quality."

We discuss drawbacks of existing methods in the paper (line 56-61, line 85, line 94, line 100). The SediNet approach by Buscombe (2019) focuses on a different application (clean sediment and sand samples) and the experimental flaw discussed above (c.f. footnote 2 in the paper) does not allow for a meaningful comparison. To our knowledge ours is the first work that evaluates grain size estimation over entire gravel

bars in river beds, and we are not aware of a comparable study regarding geographical generalization.

Due to the end-to-end learning, the proposed CNN is able to extract global features that are informative about grain size beyond the sensitivity of human photo-interpretation as well as traditional photosieving that relies on local image gradients to delineate individual grains. Even the latest work of Purinton and Bookhagen (2019) can only detect individual grains that have a b-axis 20x the ground sampling distance. Also previous statistical approaches (Carbonneau 2004) are limited by the input resolution and can only predict D50 down to 3 cm, so we do believe that our approach overcomes some of the limitations of prior art.

Moreover, the CNN is a generic learning machine and can be seen as a "Swiss army knife" for grain size characterisation from images. It can be used to directly predict the full grain size distribution at a specific location, but the same architecture can also be trained to directly predict other desired grain size metrics derived from the distribution. A further advantage over most competitors, with large practical potential, is that our method scales to entire gravel bars, resulting in detailed, high-resolution maps and more robust spatial aggregations.

Obviously, creating a large, manually labelled training dataset is time-consuming, a property our CNN shares with other supervised machine learning methods. However, at test time the proposed approach requires no parameter tuning by the user, a considerable advantage for large-scale applications, where traditional image processing pipelines struggle, since they are fairly sensitive to varying imaging conditions.

Code

The project is implemented using Keras, with Tensorflow as a backend. The code and a demo will be published together with the paper. Unfortunately, we are unable to distribute the full dataset for commercial reasons, as it is owned by a private company (who also created it at their own cost).

References

Buscombe, D. (2020). SediNet: A configurable deep learning model for mixed qualitative and quantitative optical granulometry. Earth Surface Processes and Landforms, 45(3), 638-651.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning, MIT Press, http://www.deeplearningbook.org. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

Veit, A., Wilber, M. J., and Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. In Advances in Neural Information Processing Systems (pp. 550-558).

Purinton, B. and Bookhagen, B. (2019). Introducing PebbleCounts: a grain-sizing tool for photo surveys of dynamic gravel-bed rivers. Earth Surface Dynamics, 7(3).

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

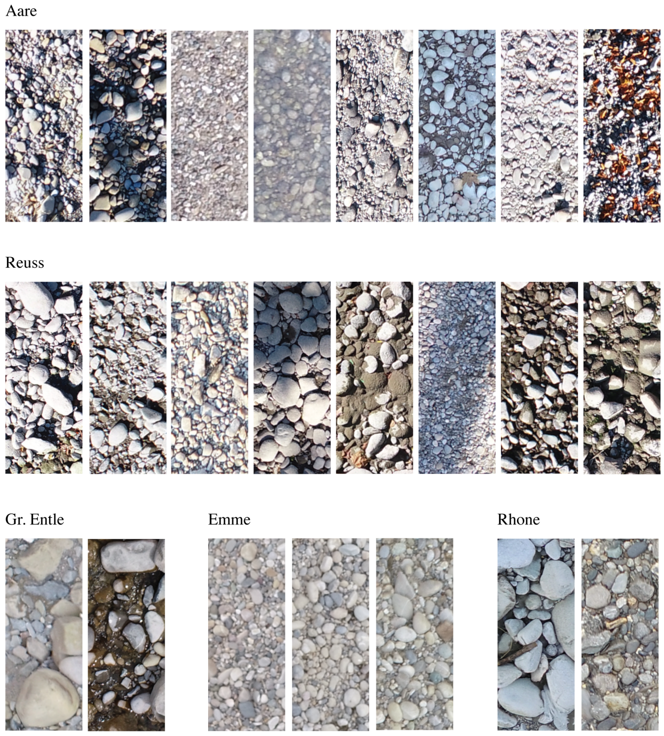

Fig. 1. Tiles grouped by river name

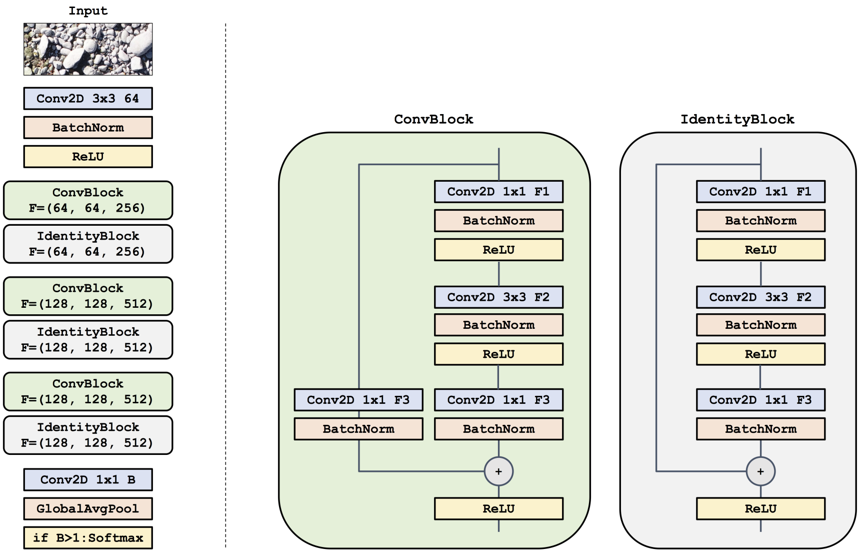

Fig. 2. Illustration of the CNN architecture

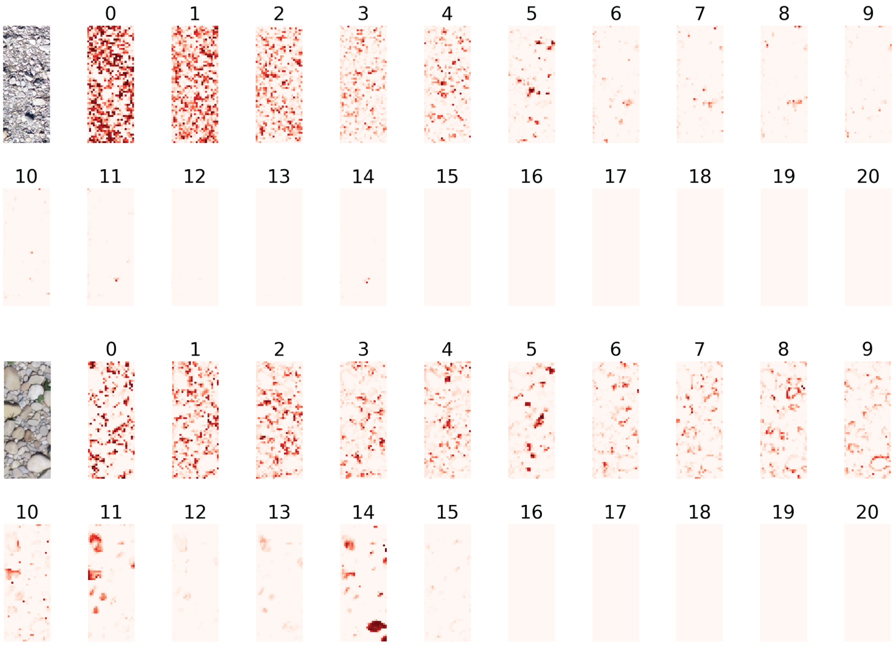

Fig. 3. Activation maps after the last convolutional layer for two examples. Each of 21 maps corresponds to a specific bin of grain sizes, where bin 0 belongs to the smallest, bin 20 to the largest grains.