Earth Surface Dynamics Discussions

Earth Surf. Dynam. Discuss., referee comment RC2
https://doi.org/10.5194/esurf-2021-67-RC2, 2021

Comment on esurf-2021-67

Anonymous Referee #2

Referee comment on "Convolutional neural networks for image-based sediment detection applied to a large terrestrial and airborne dataset" by Xingyu Chen et al., Earth Surf. Dynam. Discuss., https://doi.org/10.5194/esurf-2021-67-RC2, 2021

The authors develop a new deep learning method to determine grain size from photographs. This is done by assembling a dataset of images of grain size, manually labeling the images, and developing an analysis pipeline that includes basic computer vision as well as a fully convolutional neural network. They compare the results to manual counts, as well as an existing computer vision technique for grain size determination from pictures.

My comments are as follows:

Abstract: I would cite BASEGRAIN if it is discussed multiple times in the abstract.

L9: ‘current methods are largely based on detecting grain intersticies”. In the paper you mention many techniques that do not use this technique, so is this true?

L21: I would go with 512 pixel *512 pixel or 512 *512 pixels

L25 ‘is time consuming’

L41: I think Rubin should be mentioned here.

L70: ‘CNN techniques have proven..’

L71: what are ‘suboptimal conditioned’ images?

L72: My understanding is that the classic Ronneberger UNET is not considered ‘state of the art’ at this point, considering how fast the field is moving.

L77-L86: this section should be rewritten for clarity. I got a bit lost. What did Mueller do? And where has the UNET been applied to orthophotographs?

L119 and Table 1: First, I do not count 136 images in the training column of table 1. can you explain? Second, is this the train/validation split? Or the train+validation/test split? How much data was used for training, validation, and testing? Was there a true hold out test set that was not looked at until the model was fully trained?

L126-128: please revise for clarity. why is Ronneberger cited here when you are discussing grain size studies?

L135: 203 or 202?

L135, 138: remove word ‘masters’ from here and any other place in the ms. Replace with something else

L137: This is interesting, please describe how this worked in detail. What if operator 2 found an error? Or a missing grain? How was this dealt with?

L138-151: It is great to see some inter-rater agreement work being done here. But I am having trouble interpreting the results. One option is to use a more standard ML metric for segmentation tasks, like intersection over union. Another idea is to explain what the reader should understand from those errors: how much actual error is there between labelers? How much does it change grain size measurements? I think there needs to be a sentence to help a reader get intuition on this metric and these numbers..

L156: Please cite the software library

L158: what does ‘better overall performance’ mean, specifically?

L179: I think a paragraph is missing that described how a mask from UNET was converted to actual grain size metrics?

Section 3.2.2: Please discuss more details about the implementation: what ML framework did you use? What were the hyper parameters: Loss function, the optimizer, the learning rate, the batch size, if you chose a weight initializer, how many epochs, the callbacks/ stopping conditions, how many filters were used in the encoding and decoding layers, and how many total tunable parameters were in the model. Also what was the training/validation split, and was it random?

L203-205: please explain the calibration procedure, and each of these parameters.

Section 4.3: have these images been seen by the UNET? Are they a hold-out set?

L294-303: Please explain these errors. You might give examples.

L308-309: Please explain more, add a citation if one exists, and/or present data if you have it.

Section 5.1: The use of acronyms and symbols in this section makes it difficult to read. The same is true for other discussion/conclusion sections, with respect to the Error metrics.

L331-333: these lines makes me think that perhaps you could have gotten a better result if you optimized the parameters better, or consulting an expert on BASEGRAIN. Please discuss how you optimized the parameters or how you chose the settings you used to get the best possible results? I think it is important to do this, since ML requires many design decisions and hyper parameter tunings too.

In this part of the manuscript and in general there is a significant focus on why this method was better than basegrain specifically. I know there are many other automated grainsize detection routines, so I wonder if the manuscript can instead position itself to discuss the strengths of the GrainID method generally, instead of how good the technique is compared to BASEGRAIN. This comment is for this Line, but is a general comment for the manuscript as a whole.

Section 5.4: I appreciate that the trained unet takes 5-22 seconds to run for each sample with a GPU. However, I do think the manuscript needs to discuss how long it took to train the model. Keep in mind that BASEGRAIN and Pebble counts do not require model training. Additionally, picture methods require collection and analysis back at a lab, vs being able to count grains in the field.

L396: so these images were never seen with the model?

L410-412: Please clarify the purpose of these lines.

Section 5 generally, and L413-417 specifically: Can this section be generalized for any photographic analysis technique for grain size? Can the manuscript present any general things that others may use?

Figure 3: Can the manuscript explain a bit more what practical insight the reader should get from this figure? I don;t have much intuition for the metric.

Figure 4d: I do not think the UNET implementation presented here has any 'Crop' operations, just 'copy' (green arrow)? I think the red arrows should be ConvTranspose (or upsample), and not pool?

Table 1: Training column does not sum to 136 images for me. Also what about validation?