Vesuvius Ink Detection Followup

Scrolling the Size: Investigating Model Performance on Lower Resolution 3D images

Yannick Kirchhoff, Maximilian Rokuss, Benjamin Hamm from OverthINKingSegmenter

Introduction

In the field of machine learning based image analysis, the impact of resolution on model performance is a significant concern. While the natural image domain is facing challenges with processing or synthesising ever increasing high resolution images and the medical image domain deals with three dimensional images, often with different resolutions along each dimension, this write-up analyses a specific use case combining the two fields. Fueled by the recent Kaggle Vesuvius Challenge on Ink Detection/Segmentation [1] we perform an analysis on how the resolution of large scale three dimensional volumes affects model performance.

The prior Vesuvius Ink Detection competition revolved around the challenging task of detecting ink within 3D X-ray scans and deciphering their content. Ancient scrolls originally housed in a Roman villa near Pompeii, which was buried by a volcanic eruption almost two millennia ago were found. These scrolls, due to the volcanic heat, are now carbonized and fragile, making physical opening impossible. Alongside the scrolls a few fragments were found as well. Due to their readability using infrared lighting, labels for some of the X-ray scans could be crafted. The scrolls and fragments were scanned using a complex procedure involving a particle accelerator to capture the extremely fine structures of the paper/-papyrus. In the end a dataset consisting of 3D X-ray scans with 4µm resolution, infrared images, and hand-labeled ink masks for three papyrus fragments was assembled. Scans at 54keV and 88keV energy levels are provided for every fragment [2]. The hope was that modern technology can hold the key to unlocking their secrets. Ultimately, the competition embodies a quest to unravel hidden knowledge preserved for centuries through the integration of cutting-edge imaging and analysis techniques.

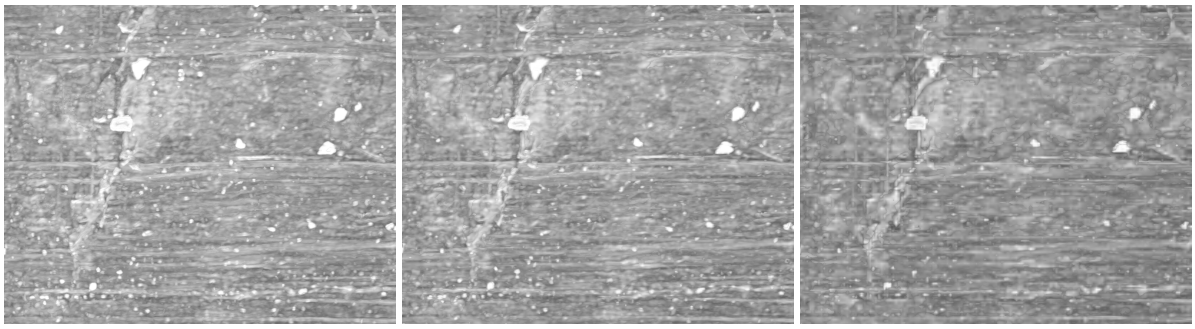
In this study, we focus on retraining and evaluating our 7th place solution [3] in the above challenge using downscaled volumes that mimic an 8 or even 16 micron resolution, as opposed to the standard 4 micron resolution, i.e. lowering the resolution by a factor of 2 or 4 respectively. Our primary objective is to assess how this difference in resolution affects model performance.

Our investigation is not just an academic exercise; it holds practical implications. We aim to understand how or if models need to adapt to function effectively with these altered resolution conditions. Moreover, we delve into the broader implications for scanning techniques since the comparison between 4, 8 and 16 micron voxel sizes could add another layer of interest. On the one hand, more accessible scanning methods that don't require complex particle accelerators might become feasible while still yielding a comparable result. On the other hand, there's a trade-off between a wider field of view and finer details. The AI model observing a larger area could gain a performance improvement from the contextual information of (connected parts of) letters, but the lower resolution means sacrificing some level of detail. Since the individual letters are rather on the order of centi- or millimeters, respectively, the FOV of the existing model ($\approx$ 2mm $\times$ 2mm) only covers a very small percentage of the whole letter (see figure 2). Finally, having a model trained on 8µm resolution will bridge the gap between the fragments and the scrolls, latter being scanned at 8µm resolution.

In essence, our study explores how resolution affects model adaptation and performance. By delving into simulated resolutions, retraining approaches, and performance analysis, we aim to uncover the fundamental role that resolution plays in computational modeling in the case of large scale three dimensional images.

Simulating lower resolution: Preprocessing of the data

In order to emulate the slices at a lower resolution they were resampled to half or one fourth of the original resolution while increasing the spacing between voxels by a factor of two or four respectively. This turned out to be a very complex and memory consuming process due to the size of the images and could only be performed on a machine with 128GB RAM. However, this preprocessing approach was chosen over just resampling input patches on-the-fly or introducing an average pooling operation as the first layer in order to leave the original pipeline untouched enabling a fair comparison. The size of the x and y dimension varied per scroll, but the z dimension of the three resolutions resulted in $z_{4\mu m} = 65$, $z_{8\mu m} = 32$ and $z_{16\mu m} = 16$ respectively. This also meant that the preprocessing step of the original OverthINKingSegmenter [3] had to be adapted accordingly, since it is tuned by default to extract 32 z-slices determined by an intensity analysis. Therefore, only the 16 or 8 most "informative" slices were extracted. Fig. 1 shows a zoomed in patch from the first scroll taken from a 3D visualization at all three resolutions.



((a)) 4µm ((b)) 8µm ((c)) 16µm

Figure 1: Visualization of a patch from the first scroll resampled to the three different resolutions. The deteriorating level of detail and clarity of fine structures is clearly visible.

Experiments

To fully assess the hypothesised performance improvement of a larger field of view (FOV) our conducted experiments with altering the resolution were two fold.

On the one hand we retrain our exact pipeline from the challenge submission with images resampled to lower resolutions, as described above, mimicking the images being taken with this resolution in the first place. Consequently, this results in a larger FOV of the network since the input size (also called patch size) is not altered but now contains two or four times more physical area. On the other hand we also assess the networks performance when just the resolution is altered but not the field of view, i.e. the area the network actually "sees". Therefore, we not only downsample the images but also decrease the patch size of the network by the same factor as well. The field of view in z-direction is always the same, as we employ the same slice selection as in our original solution. The experiments with the same field of view as the original contribution at 4µm resolution are denoted by “FOV” in the following.

Table 1 and Fig. 3 show quantitative and qualitative results of the trainings with different resolutions and field of views on the fourth fragment, which was also used as test fragment in the original challenge. Surprisingly, for both downsampled resolutions of 8µm and 16µm the configurations with a

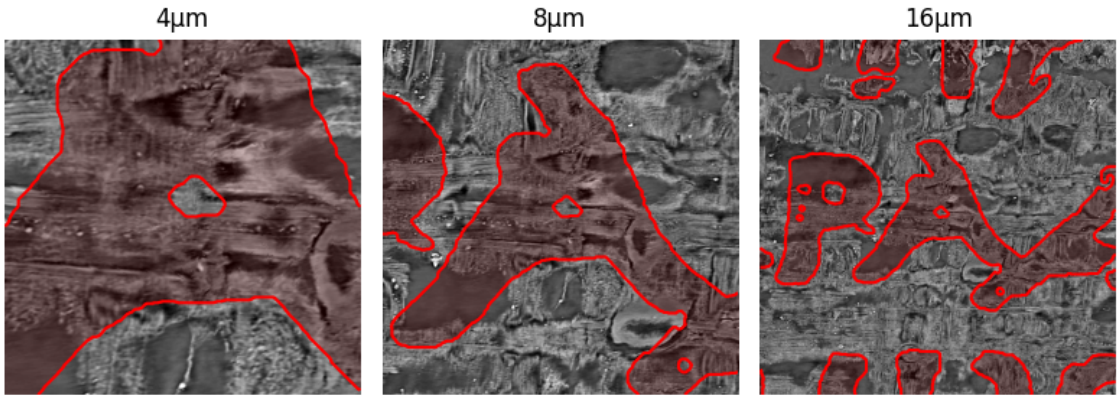


Figure 2: Patch of size 512×512 at different resolutions

restricted FOV perform significantly better than the networks, which get more context. This can probably - at least partially - be explained by the splitting of the data we performed for our solution and also applied for this evaluation. The individual fragments were split into 25 pieces each, where every piece contains approximately the same amount of foreground pixels. This leads to several pieces which are only slightly larger than the patch size of 512 in the original 4µm resolution. For the downsampled scans these pieces are zero-padded during training, which increases the amount of background pixels. This gives an additional weight for the background in the cross entropy loss. As we wanted a direct comparison with the same data splits and training setups, we did not change that for this evaluation.

The training with a fixed field of view at 8µm performs only slightly worse on the test fragment compared to the 4µm training, indicating that most features the network uses are still visible for the network. Qualitatively, the changes mostly don't affect the readability of the predicted letters. At 16µm the drop in performance is more significant, both in the quantitative, as well as the qualitative assessment. However, it should be noted here that the network architecture was slightly modified to incorporate the patch size of only 128×128, which might also have affected the model's performance.

	TP	FP	FN	Sorensen DSC
4µm	2140677	729988	2469604	0.6651
8µm FOV	510675	182347	641592	0.6506
8µm	544798	251820	607469	0.6278
16µm FOV	109302	53280	178830	0.5823
16µm	117145	98252	170987	0.5095

Table 1: Quantitative results of different configurations on fragment 4

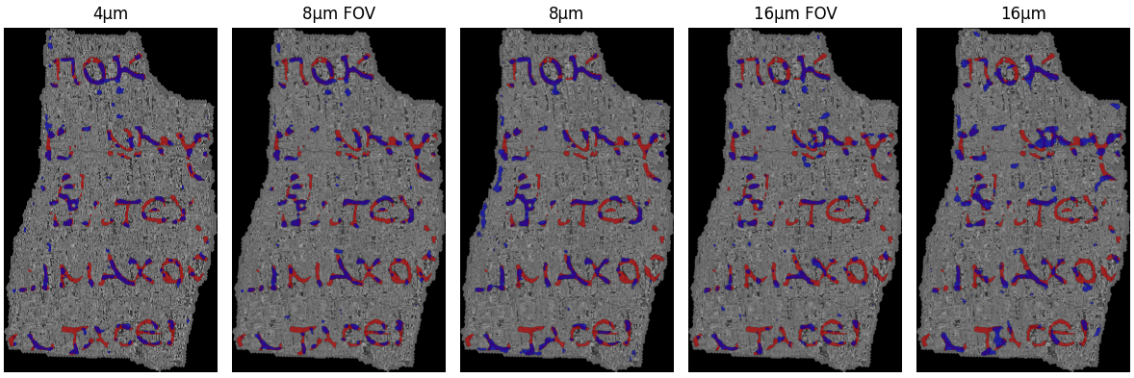


Figure 3: Qualitative results of different configurations on fragment 4. Ground truth is shown in red, model predictions are shown in blue.

Conclusion and outlook

For this evaluation we retrained our contribution to the Vesuvius challenge on downsampled resolutions of 8µm and 16µm, respectively. Keeping the field of view constant, we found that there is only a slight performance drop going from 4µm to 8µm and a larger drop in performance going to 16µm resolution. Enlarging the patch size and therefore giving the network more context performed worse than the fixed field of view setups, which is probably due to the splitting of the training data.
With a different splitting technique, guaranteeing for large enough pieces for all patch sizes, we expect increased performances with respect to the fixed field of view. However, this would require retraining all configurations on all folds to achieve comparability between the different resolutions. This was out-of-scope for this evaluation. For future evaluations it might be sensible to include the fourth fragment into the training.

References

[1] Alex Lourenco, Brent Seales, Christy Chapman, Daniel Havir, Ian Janicki, JP Posma, Nat Friedman, Ryan Holbrook, Seth P., Stephen Parsons, and Will Cukierski. Vesuvius challenge - ink detection, 2023.

[2] Stephen Parsons, C. Seth Parker, Christy Chapman, Mami Hayashida, and W. Brent Seales. Educelab-scrolls: Verifiable recovery of text from herculaneum papyri using x-ray ct, 2023.

[3] Yannick Kirchhoff, Benjamin Hamm, and Maximilian Rokuss. OverthINKingSegmenter: 7th place solution for the vesuvius kaggle challenge, 2023.