Nucleus segmentaion of DNA comets as preprocessing to DNA damage assessment.

Danylo Kolinko, Ukrainian Catholic University

Abstract—**Comet assay is a procedure for assessment of DNA damage for eukaryotic cells. This technique is used in numerous researches in biology and medicine. For instance, it is used to test preparation for age-associated diseases: atherosclerosis and cardiovascular disease, cancer, arthritis, cataracts. Despite its importance, this procedure requires lots of human supervision. Humans should identify every single cell and exclude all overlapped instances. To put into perspective, a single frame can contain more than thousand of comets. So there is some space to automate this procedure. The largest pain in the ass is the localization of cells. So cell segmentation would be a good idea, and it is the goal of this project. I aim to perform classic unsupervised segmentation techniques as well as deep learning segmentation for effectiveness comparison.**

Index Terms—**nucleus segmentation, deep learning, DNA damage, comet assay, etc.**

Figure 1. Comet assay example.

I. INTRODUCTION

COMET assay is the techinque that essentially aims to decompose DNA in the individual cells. Short description of this procedure goes as follows:

- Cells are dispersed in agarose, and this mono-suspension is cast on a microscope slide.

- These cells are exposed to a subject of testing (may be some preparation or radiation).

- Then cells are subjected to lysing (particular alkaline-based procedure, which removes all non-DNA structures from a cell: RNA, membrane, etc.).

- Then cells are submerged into different alkaline and exposed to electrolysis. Since undamaged DNA has a solid structure, there is no deformation, but damaged DNA starts to float towards the anode. This way we can decompose a cell into damaged part (tail) and undamaged (head).

- The comet assay than can asses a level of DNA damage by intensity and size of tail and head. And the new deformed object is called a comet.
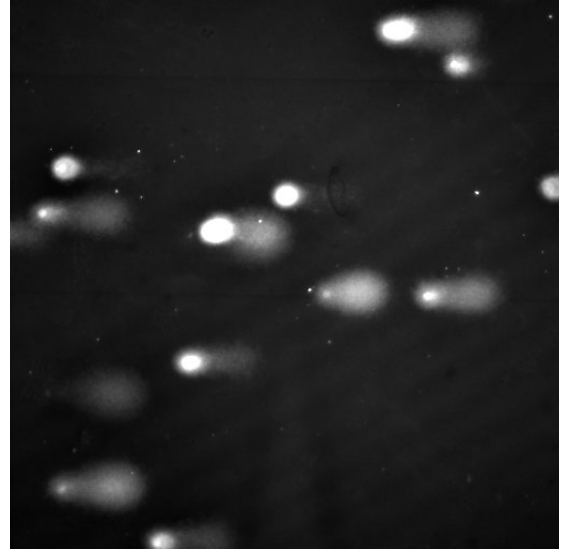
Here may be observed just a small piece of the comet's frame with 14 comets. Nowadays, researcher supposed to localize each cell, which is time consuming and inefficient. With the rise of deep learning, we have got new powerful tools for segmentation of complex objects. Moreover, progress in an application of CV in biology research has shown some of the most impressive results. However, apparently in the comet segmentation there is no prior research using deep learning architectures.

The goal is too test UNet architecture with its modifications, run several experiments and learning strategies and compare the outcome to unsupervised classic segmentation technique.

II. DATASET AND PREPROCESSING

As I found out there is limited open-source data on this procedure, which might be the reason on absence of prior research using DL. Initial dataset contained 70 frames of comet assay. Additionally, 120 frames was found on the web. Then all this data was manually unlabeled resulting in 190 frames with 2.7 thousands of comets.

Basically, there are 7 classes of comets in terms of its image representation. One of them is scoreble and the rest is different reasons of unsuitability for assessment. I segment all of them and later on will classify. That is because now we have ways to refine those classes and assess them as well.

III. RELATED WORKS

There are few developments on automated comet assay. For example, tool called CellProfiler offers a great of different pipelines for cell processing, comet assay as well. This instrument has shown poor results testing on noisy and complex frames.

Other tool HiComet offers no working tool to test, but has thorough description of its segmentation method, which is watershed preprocessed with distance and wavelet transforms. Although, authors do not provide any metrics of their segmentation. This method I chose to re-implement and obtain relevant metrics.

Regarding deep learning segmentation I was guided by 2 papers:

- Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images (Caicedo et al. 2019)

- Segmenting nuclei in brightfield images with neural networks (Fishman et al. 2019)

IV. UNSUPERVISED PIPELINE

This preprocessing steps are simple:

- Smoothing as preprocessing step to reduce noise

- Binarization which was performed as threshold on the first valley point on the image histogram. This way the difference in elumination could be taken into account (unlike classic Otsu threshold).

Then filtering and overlap correction steps are performed:

- Morphological opening operator is applied to seed out some noise.

- Distance transform is performed coupling with wavelet transform as overlap correction markers generation steps.

- With mentioned markers watershed segmentation was applied

After this step Fourier shape descriptors were obtained which were expected to have dominant low-order frequencies and others were seeded out.

Described pipeline I implemented using general opencv tools. This pipeline was then tested on the collected dataset and its segmentation masks as ground truth with jaccard similarity and dice similarity.

Resulting in mean jaccard = 0.15 and mean dice = 0.22 with high variance; this strategy was able to obtain 0.65 dice

for simple images, but failed for complex ones.

Having this method as the baseline I moved forward to explore deep learning techniques.

V. DEEP LEARNING SEGMENTATION

In both reviewed papers on cell segmentation, which were mentioned above, was shown dominance of UNet architecture proposed by (Ronneberg et al. 2015). This architecture is fairly simple: encoder and decoder with skip connections and cropping. Although this architecture was proposed for bio-images developed in mind with its challenges like small datasets, etc., this architecture achieved stunning results in absolutely different segmentation problems.
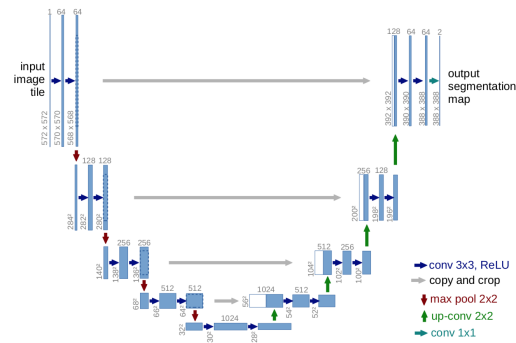

Figure 2. Unet structure.

That is much more convinient to describe a path of different techniques that I tried to improve segmentation performance.

Well, as a starting point I built UNet11 and UNet16. Among this two using cross entropy loss, unet16 was harder to train, but it achieved better results.

Later on I used pretrained vgg encoders, which obviously reduced training time.

Another challenge was to increase confidence in prediction; for this purpose was used different training strategies.

First one - to use dice loss as loss function which was confident enough, but did not properly learn the shape. So dice loss was tried as second loss function after 30 epochs of binary cross validation (BCE from now on) and different approach was tried as well- weighting them both and using as loss function. As well jaccard loss was later tried, yielding similar results.

Either of these did better than previous strategies.

Here I should mention that I used different augmentation to enhance preformence : rotation, flip, trasposing, blur, noise, crop, shift and brightness as well as grid distortion.

Nevertheless, major issue was artifacts and debris on the frame. Since there were just few non-comet object with high brightness, network was hardly able to distinguish comet

from light debris. The solution was to generate more light random blobs and force to learn other comet patterns rather than brightness.

I tried also to manually adjust learning rate which yielded another 0.04 points to dice score.

That is final improvement for now. Here was achieved jaccard (IoU) score of 0.3 and dice score of 0.46, which is quite far from optimal but was able to yield convincing results.
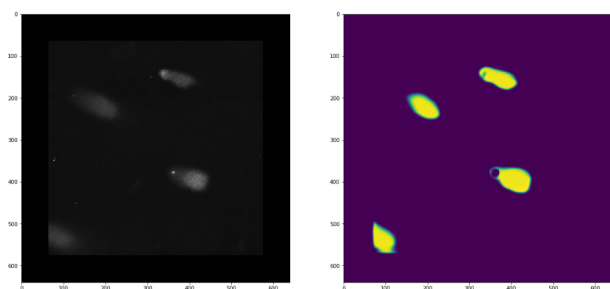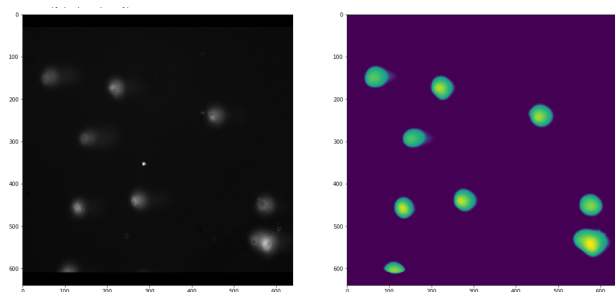

Figure 3. Segmentation result

Figure 4. Segmentation result

Other than that ResNet was also tried as encoder for UNet but didn't yield better results.

One future development that I didn't have time to explore is penalized border segmentation, used to split the comets better and add weight to tail recognition.

VI. CONCLUSIONS

Unsupervised segmentation has its limits in robustness. Supervised in data volume. Work on this project is still in progress; So there can not be any final conclusions. There are plenty ways to guide your network towards good result and many of them are yet to be explored. Lot's of things did not work out that I described, but plenty did work as expected. So I'll keep exploring.

REFERENCES

[1] CellProfiler, cell image analysis software
https://cellprofiler.org/

[2] HiComet
https://bmcbioinformatics.biomedcentral.com/art

[3] Juan C. Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W. Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J. Theis, Anne E. Carpenter, "Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images", 2019
https://onlinelibrary.wiley.com/doi/10.1002/cytc

[4] Dmytro Fishman, Sten-Oliver Salumaa, Daniel Majoral, Samantha Peel, Jan Wildenhain, Alexander Schreiner, Kaupo Palo, Leopold Parts, "Segmenting nuclei in brightfield images with neural networks", 2019
https://www.biorxiv.org/content/10.1101/764894v1

[5] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", 2015