Original Research Article

POLYOMAVIRUS LOCALIZED WITH DEEP LEARNING METHODS

.

ABSTRACT

A high-resolution 3D reconstruction of virus particles or other macromolecular structures needs thousands of particles, often selected from transmission electron microscopy images; the ideal number is utopic, but the recent advances in deep-learning methods makes the reach of a comfortable number attainable with a high level of confidence. The main difficulties consist in the confusing background where a mix of artifacts introduced by the negative staining method coexist with biological debris, and low contrast inherent to the acquisition process of the samples. Recent algorithms based on convolutional neural networks can overcome the major issues and process thousands of images in a short amount of time, as a few seconds are required for processing each image, localizing all the particles of interest. We applied one of such algorithms to a dataset of polyomavirus images containing about 390 particles and evaluate its performance against human curation - 88.9% of the dataset were processed with zero false positive detections.

Keywords: Polyomavirus, Particle picking, 3D reconstruction, Deep-learning methods

1. INTRODUCTION

The selection of biological particles in electron microscopy images with efficiency better than 75% compared to a human expert is the main goal in particle detection for high resolution three-dimensional (3D) reconstruction [1]. Considering that more than 100.000 particles are needed to achieve a resolution better than 1 nm, it is a time and cost expensive procedure in human labor [2], as only microbiologists with high level of expertise can be involved and is a subjective work, although interactive methods can solve some situations [3]. It could be viewed as a problem of image segmentation, where there is only one class of objects of interest (OoI), and has been subjected to many attempts of automation always including human intervention at some point - at the initial stage to build the templates [2], [4]-[6], or in the final stage, to decide the threshold applied to split the set of candidates [5] in accepted/non accepted.

The problem as we state it aims to detect all the biological particles of one type, polyomavirus in this case study, presented in a transmission electron microscopy (TEM) image, as long as the particle would be recognized by an expert observing the sample for diagnostic purposes. Negative staining TEM images have a considerable dynamic range and high signal to noise ratio (SNR), but the resolution is limited by the sample's characteristics. TEM importance remains worldwide present [7]-[9], as electron microscopes are commonly found in laboratories and hospitals and continue to be a catch-all method able to identify new and/or unexpected viruses. In emergency situations such as outbreaks or bioterrorism attacks [10], the detection of new or unknown pathogens is more likely to come from such methodology largely available than from molecular and immunological methods much more expensive that must be targeted to specific pathogens.

Amongst publications concerning location of virus particles from TEM images, we could find a few using machine learning techniques [11] and neural networks [12] since the beginning of the boom of neural networks, and more recently applied to feline calicivirus particles [13] and human cytomegalovirus [14].

2. MATERIAL AND METHODS

The images in the data set come from a collection of TEM micrographs used for diagnostic and monitoring purposes of human infection by polyomavirus, from samples of urine from bone marrow transplanted patients. Urine samples were centrifuged at 3,000 rpm for 20 min, and the cleared urine was ultracentrifuged at 29,000 rpm for 75. The pellet was suspended in a drop of distilled water. A drop of the sample was placed on a formvar-coated copper grid for 1 min and washed with seven drops of 2% ammonium molibdate. Excess fluid was removed by touching with a strip of filter paper. Grids were observed in a CM10 Philips electron microscope. The screening was done at a magnification of 46,000. The presence of positive cases of polyomavirus was recorded at a magnification of 52,000 with Electron Microscope Film 4489 (Eastman Kodak, Rochester, NY, USA), and then processed with diluted D19 developer Kodak for 4 min at 20ºC. Negatives obtained from virus-positive cases were scanned with an Epson Perfection 4990 Photo Scanner in gray scale at 1,200 dpi and 16 bits. All images in the data set have a pixel size of 0.41 nm and were pre-processed with a contrast-limited adaptive histogram equalization in Matllab environment.

The deep-learning method we choose to tackle the problem is open-source software, implemented in Pytorch. We use one of the latest developments in one stage algorithms based on convolutional neural networks (CNNs) since the introduction of the concept You Only Look Once (YOLO) in 2016 [15]. The 5th version, YOLO v5 was made publicly available in a GitHub repository [16] in 2020, with the advantage that the algorithm can be trained in any exogenous data set.

The train stage consists in annotate a considerable number of images, boxing all occurrences of the OoI to be detected. Tools available online such as [17] allow to upload a data set, annotate it according to our goals, and download a set of output files with the annotations in a user defined format. Another advantage is the possibility of transfer learning – since many basic features are common to all detection problems (edges, contrasts, forms, etc.) an already heavily trained network can be used to implement a new problem. The new discriminators will define the last layers of the CNN, tunning the detector according to the details of the specific problem.

This kind of algorithm must be trained for one specific virus particle, in this case, the polyomavirus; the possibility of data augmentation enables the detection with slight differences in working magnifications and radiometric characteristics but cannot be applied to a different virus that present different texture, structure and typical form, which means different attributes. Data augmentation is the introduction of small randomized changes to train data, in both forms: geometric changes in dimensions, translation, rotation, shearing, and radiometric changes in color components and intensity. This is very appealing for particle picking because whatever number of train images we use, it’s very unlikely that all possible positions, sizes, and tonalities are included in the train set.

YOLO v5 performance is evaluated using the concepts of precision and recall; precision is the accuracy of the network, which refers to the proportion of targets correctly classified by the network (the ratio true positives/all that has been classified as positive, meaning true plus false positives). Recall refers to the proportion of the targets correctly detected (the ratio true positives/all that is positive, meaning true positives plus false negatives).

We use the model x of YOLO v5 with the default hyperparameters and annotate about 30% of 54 sub images of 1024x1024 pixels with one class of objects. Eleven images were used for train and seven for validation. The weights resulting from the train can now be used to inference in any other image with the same characteristics to identify and localize the OoI. Although the train stage can take a few hours (18.4 h for 500 iterations in a laptop equipped with processor dual Core Intel i7-10750H, 16 GB SDRAM and a NVIDIA GeForce RTX 2060), the inference takes around 2500 ms for each micrograph.

3. RESULTS AND DISCUSSION

The detection of polyomavirus particles in the test dataset correctly identifies all the instances of the virus in 86.1% of the test images. With a confidence threshold of 0.65, the overall precision achieved in the dataset is 96.1% and the recall attains 94.8%. A typical image with an heterogenous background (Figure 1) shows how the algorithm performs, detecting even partial particles with high confidence level.

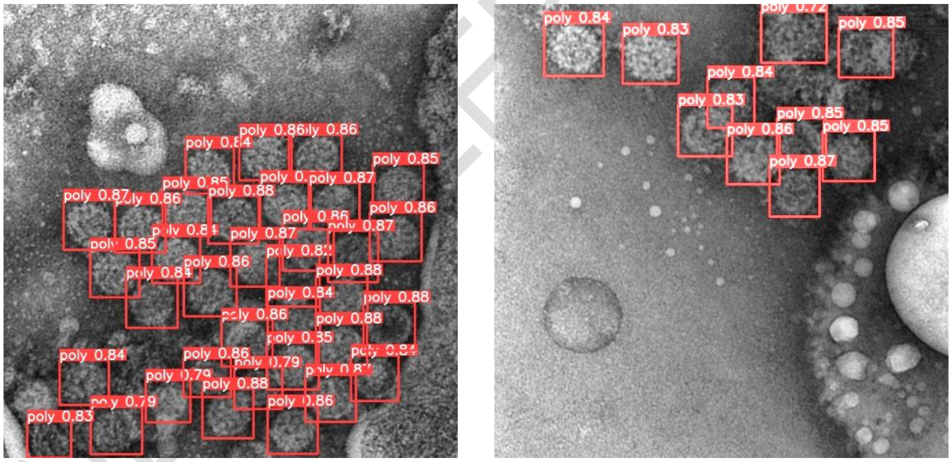
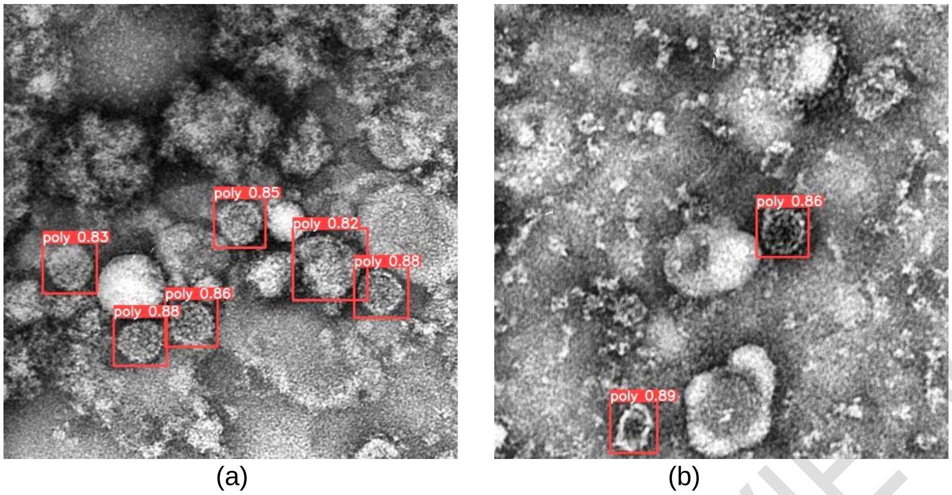


Fig. 1. Two examples of polyomavirus localization with YOLO v5. The number associate to each detection is the probability that the object detected is an instance of the objects of interest. The confidence threshold is a parameter applied for inference, allowing to eliminate detections with lower confidence; the value used in this data set is 0.65.

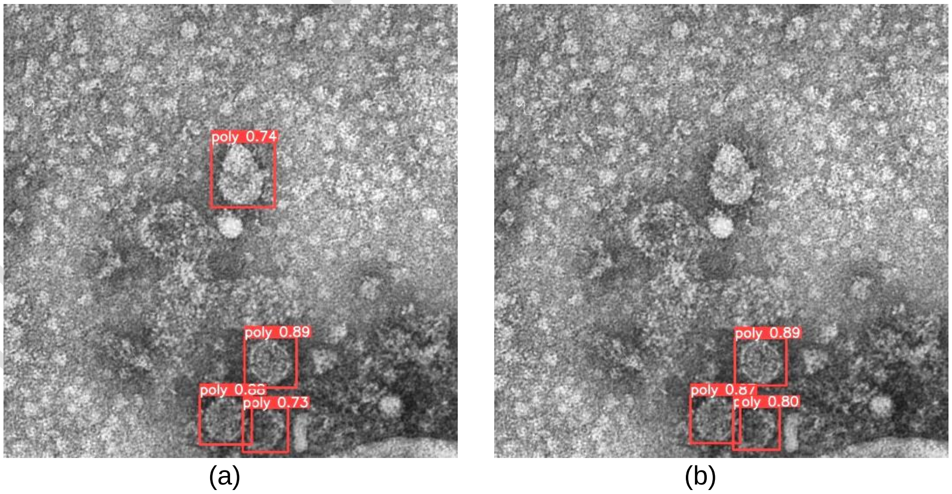
The few misclassifications can be often understood considering the characteristics of the areas erroneously considered as particles of interest (Figure 2).



(a) (b)

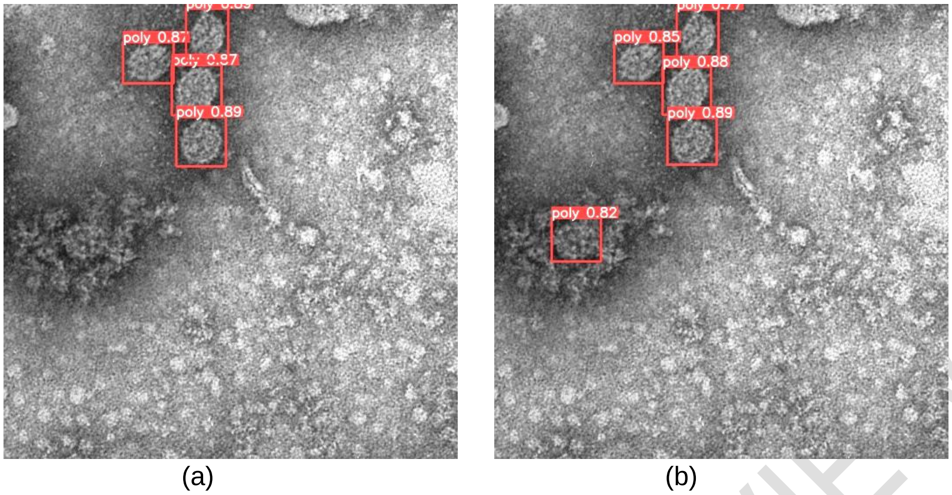
Fig. 2. These images contain one false positive each: in (**a**) the second detection from right, although with high confidence (0.82), is a wrong detection; in (**b**) the detection near the bottom of the image is a false positive.

In the dataset of polyomavirus available, more than 86% of the images were processed without any false negative, and nearly 78% with zero false positives. Analyzing the few misclassifications, it looks like small changes in data augmentation parameters at train stage could eventually lead to better results, as some false positives could be the result of a 50% change admitted to the scale in the hyperparameters settings. The procedure was repeated with zero percent change in scale in the data augmentation parametrization, leading to an increase in precision to 98.0%, while recall decrease to 87.1%. While some obviously false detections were eliminated (Figure 3), it contributes to identify almost indiscernible particles that were not considered in human curation but could actually be OoI (Figure 4).



(a) (b)

Fig. 3. The inference with the first train (a) has a false positive that is eliminated (b) when the weights issued from the second train are used. The train setting differs for one hyperparameter, scale set to 0.5 in the first train and to zero in the second.



(a) (b)

Fig. 4. In this image the four particles present were correctly identified with the first parametrization (a), but the second train (b) contributes to a reassessment of human curation, putting in evidence a particle considered a false positive, that could actually be an object of interest.

The inference with the weights from the new train process 88.9% of images with zero false positive detections against 77.8% with the previous train, suggesting this could be a better option, although some good particles should be missed, as only 69.1% of the images has zero false negatives against 86.1% with previous train.

4. CONCLUSION

Deep learning methods are a recent tool with broad applications, and we intended to show the potentialities of this algorithm (YOLO v5) applied to polyomavirus picking, and the difficulties inherent to the process. Future work in larger datasets and different particles would help to establish an ideal parametrization in line with the goals, but this seems a promise line of work to pick particles from large datasets of micrographs in order to attain numbers that can ensure a good quality for 3D reconstruction with high efficiency and reduced time and human labor costs.

COMPETING INTERESTS DISCLAIMER:

Authors have declared that no competing interests exist. The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

REFERENCES

1. Glaeser RM. Historical background: Why is it important to improve automated particle selection methods? J. Struct. Biol. 2004**;** 145(1–2):15–18.
2. Wong HC, Chen J, Mouche F, Rouiller I, Bern M. Model-based particle picking for cryo-electron microscopy. J. Struct. Biol. 2004;145(1–2):157–167.
3. Shah AK, Stewart PL. QVIEW: Software for rapid selection of particles from digital electron micrographs. J. Struct. Biol. 1998;123(1):17–21.
4. Roseman AM. FindEM—a fast, efficient program for automatic selection of particles from electron micrographs. J. Struct. Biol. 2004;145(1–2):91–99.
5. Hall RJ, Patwardhan A. A two-step approach for semi-automated particle selection from low contrast cryo-electron micrographs. J. Struct. Biol. 2004;145(1–2):19–28.
6. Short JM. SLEUTH—a fast computer program for automatically detecting particles in electron microscope images. J. Struct. Biol. 2004;145(1–2):100–110.
7. Goldsmith CS, Miller SE. Modern uses of electron microscopy for detection of viruses. Clin. Microbiol. Rev. 2009;22(4):552–563.
8. Correia FF, Matos B, Moura Nunes JF, Alves de Matos AP. Applications of transmission electron microscopy to virus detection and identification. Microscopy, Science, Technology, Applications and Education, A. Méndez-Vilas and J. Diaz, Eds. Badajoz, Spain: Formatex; 2010.
9. Ong H, Chandran V. Identification of gastroenteric viruses by electron microscopy using higher order spectral features. J. Clin. Virol. 2005;34(3):195–206.
10. Hazelton PR, Gelderblom HR. Electron microscopy for rapid diagnosis of infectious agents in emergent situations. Emerg. Infect. Diseases. 2003;9(3):294–303.
11. Sorzano COS, Recarte E, Alcorlo M, Bilbao-Castro JR, SanMartín C, Marabini RJM. Automatic particle selection from electron micrographs using machine learning techniques. *J.* Struct. Biol. 2009;167(3):252–260.
12. Ogura T, Sato C. Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: A new reference free method for single-particle analysis. J. Struct. Biol. 2004;145(1–2):63–75
13. Ito E, Sato T, Sano D, Utagawa E, Kato T. Virus Particle Detection by Convolutional Neural Network in Transmission Electron Microscopy Images. Food Environ. Virol. 2018; 10:201–208. https://doi.org/10.1007/s12560-018-9335-7
14. Devan KS, Walther P, von Einem J, Ropinsky T, Kestler HA, Read C. Detection of herpesvirus capsids in transmission electron microscopy images using transfer learning. Histochem Cell. Biol. 2019;151:101–114. https://doi.org/10.1007/s00418-018-1759-5
15. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition. 2016:779-788, 1506.02640.pdf (arxiv.org)
16. GitHub. n.d. Available online: GitHub - ultralytics/yolov5: YOLOv5 🚀 in PyTorch > ONNX > CoreML > TFLite. Accessed 16 December 2021.
17. Makesense n.d. Available online: Make Sense. Accessed 22 December 2021.