A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology

Abstract:

Nuclear segmentation in digital microscopic tissue images can enable extraction of high-quality features for nuclear morphometrics and other analysis in computational pathology. Conventional image processing techniques, such as Otsu thresholding and watershed segmentation, do not work effectively on challenging cases, such as chromatin-sparse and crowded nuclei. In contrast, machine learning-based segmentation can generalize across various nuclear appearances. However, training machine learning algorithms requires data sets of images, in which a vast number of nuclei have been annotated. Publicly accessible and annotated data sets, along with widely agreed upon metrics to compare techniques, have catalyzed tremendous innovation and progress on other image classification problems, particularly in object recognition. Inspired by their success, we introduce a large publicly accessible data set of hematoxylin and eosin (H&E)-stained tissue images with more than 21000 painstakingly annotated nuclear boundaries, whose quality was validated by a medical doctor. Because our data set is taken from multiple hospitals and includes a diversity of nuclear appearances from several patients, disease states, and organs, techniques trained on it are likely to generalize well and work right out-of-the-box on other H&E-stained images. We also propose a new metric to evaluate nuclear segmentation results that penalizes object- and pixel-level errors in a unified manner, unlike previous metrics that penalize only one type of error. We also propose a segmentation technique based on deep learning that lays a special emphasis on identifying the nuclear boundaries, including those between the touching or overlapping nuclei, and works well on a diverse set of test images.