RESULTS OF AUTOMATICALLY CREATING AN ANNOTATED DATASET

Kovalenko A. S., Severin V. P.
National Technical University «Kharkiv Polytechnic Institute», Kharkiv

The study investigates a methodology for the automated generation of annotated datasets from microscopic images of biological objects. The proposed approach is designed to facilitate the development of training data for machine learning models in the field of computer vision. The proposed method builds upon the CRISP-DM framework and comprises six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These stages have been adapted to address the specific requirements of computer vision tasks. [1].

During the modeling stage, the method utilizes the k-means clustering algorithm to categorize biological objects based on their normalized area and average color characteristics. Unlike supervised methods, this unsupervised approach does not require pre-labeled data, making it highly effective in scenarios where annotations are unavailable. Microscopic images undergo preprocessing, including noise removal and color space transformation, followed by contour detection and feature extraction.

The pipeline was tested on blood smear images to detect and classify white blood cells (WBC), red blood cells (RBC), and platelets. A dataframe was generated containing object coordinates, types, and area measurements. Visualization of the results was implemented using Python libraries such as Pandas, OpenCV, and Matplotlib. Manual verification of 158 images is summarized using accuracy, precision, recall, and F1-score metrics: WBC achieved high performance across all metrics, with an accuracy of 0.9814, perfect precision (1.0), recall of 0.9814, and an F1-score of 0.9906, indicating a highly reliable classification; platelets showed strong results as well, with an accuracy of 0.8723, precision of 0.9767, recall of 0.8909, and an F1-score of 0.9318, reflecting good overall balance between precision and recall; RBC demonstrated relatively lower performance, with an accuracy of 0.7895, precision of 0.8392, high recall of 0.9302, and an F1-score of 0.8824, suggesting that while the model effectively detects most RBCs, some misclassifications still occur.

This approach offers several advantages compared to traditional manual methods of generating annotated datasets. Firstly, it is fully automated, enhancing both efficiency and scalability. Secondly, it is adaptable for use with various types of biological objects. Lastly, it considerably reduces the time required to produce an annotated dataset. The final output is a CSV file containing object annotations, which can be used to apply bounding boxes for training neural networks.

References (translated):

1. S. Kovalenko, S. Kovalenko, A. Kutsenko, M. Godlevskyi, V. Severin, A. Kovalenko, Methodology for Creating Annotated Datasets of Biological Objects in Microscopic Images, 2024 IEEE 5th KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2024, pp. 1-6,