УДК 004.85 DOI: https://doi.org/10.30837/IYF.IIS.2024.077

CELL TRACKING IMPROVED THROUGH MACHINE LEARNING

Kupriianov S.
Scientific adviser – Doctor of technical sciences, Professor Avrunin O.
Kharkiv National University of Radio Electronics, dpt. AI
Kharkiv, Ukraine
e-mail: serhii.kupriianov@nure.ua

The work is devoted to assessing the challenges and advancements in microscopy data analysis, focusing on cell tracking, object detection, lineage reconstruction, and the application of machine learning. It highlights the difficulties in accurate lineage reconstruction, the importance of metrics in assessing tracking algorithms, and the role of machine learning in proposing hypotheses and understanding complex biological systems.

Deciphering how cells self-organize is crucial in biology. Single-cell biology, illuminating processes from development to diseases, relies on computational algorithms for spatial and temporal cell biology quantification. Light-sheet microscopy's recent advancements, including multiple biochemical reporters like B.Yang, A.R. Jamieson, and others, produce volumetric time-lapse data with rich experimental imagery, surpassing current data extraction capabilities.

In recent years, machine learning (ML) has transformed microscopy data analysis, improving tasks like cell segmentation and denoising. Despite advancements in cell tracking, accurately capturing multi-generational lineages remains a challenge [1]. Current approaches use a tracking-by-detection paradigm, detecting and linking cells over time.

With the advent of convolutional neural networks (CNNs), the cell detection step has seen significant progress in recent years. Generalized segmentation and detection algorithms such as U-Net, Mask R-CNN, YOLO, Segment Anything [2] or more specialized cell-specific algorithms such as DeepCell, Cellpose and StarDi st are now able to detect cells with great accuracy, even in complex multidimensional data.

Unlike tracking-by-detection algorithms relying on heuristics, this approach prioritizes quantifying cell biology over precise tracking accuracy. It lacks flexibility with new data but posits that machine learning advancements enable learning cell behavior models by framing tracking as a learnable task [3].

Cell tracking necessitates two types of annotations: spatial-temporal marking of individual cells and detailing their temporal linking. Manual annotation acquisition is labor-intensive, with reports of weeks to years spent annotating datasets for training. For 3D + t datasets, annotations are scarcer, often providing a sparse "gold standard". Benchmarking against this limited gold standard may overlook improvements in diverse cellular behaviors.

To enhance annotated data, crowd-sourcing annotations are on the rise, emphasizing the importance of active label cleaning for improved dataset quality. High-quality datasets increase through efforts like the Cell Tracking Challenge (CTC) and the Multiple Object Tracking (MOT) benchmark, capturing varied cell types and contributing to repositories.

The tracking data, represented as a directed acyclic graph (DAG), forms G hypothesis = ⟨V, E⟩, capturing cell division events. The algorithm seeks G solution ⊂ G hypothesis to minimize tracking errors and depict cell motion, mitosis, and apoptosis, posing challenges in detection linking and lineage reconstruction.

In the tracking-by-detection paradigm, a greedy assignment strategy utilizes a cost matrix (C) for edges between vertices at consecutive time points. Solving the Linear Assignment Problem (LAP) involves algorithms like Hungarian or Jonker-Volgenant, but their O (n3) time complexity hinders large datasets.

Constructing the vital cost matrix C initially considers spatial distance, later incorporating advanced functions like Kalman filters or local flow. Despite improvements, there's room for further feature inclusion in C. This falls under local tracking, spatially global but not temporally. In contrast, global tracking considers the entire hypothesis graph while identifying the optimal set of edges.

In long-movie cell tracking, accurate lineage reconstruction is challenging [4]. It relies on precise object detection and hypothesis graph construction for mitosis events, making it more error-prone than reconstructing single-cell tracks. Measuring tracking errors is crucial, with metrics playing a key role in machine learning training loops. Common errors include inaccuracies in object detection, leading to impactful false negatives and positives, influencing G hypothesis construction and lineage tree accuracy. Several metrics assess cell tracking algorithms, with lineage-specific metrics like "branching correctness" (MBC) and "leaf retrieval score" (LRS, benchmarked at 0.75) providing user-intuitive insights into tracking performance. These metrics are implemented in open-source packages like Traccuracy.

Tracking 10 cells in a 10-frame movie yields an impractical 10^9 potential solutions in the hypothesis graph (G hypothesis) for larger datasets. Machine learning treats tracking as a learnable task with representation and hypothesis graph components. Recent discrete optimization methods include the Viterbi algorithm and Linear Programming (LP). The widely used Integer LP (ILP) relies on user-provided rules, leading to exponential complexity in novel datasets, indicating incorrect assumptions.

The goal of future ML systems is to intelligently propose hypotheses, reducing the search space rather than enumerating all possibilities. Self-supervised methods in recent studies show potential for predicting cellular events from image data. ML-enhanced tracking algorithms could leverage these predictions, and fusion of deep learning with combinatorial solvers offers an

end-to-end tracking pipeline. Our research aims to understand cell behavior in complex biological systems using ML as a framework to learn aspects of it [4]. Efforts in computer vision explore end-to-end ML tracking algorithms, like global tracking transformers, but they require large training datasets [2] and currently overlook important hypotheses like branching events. Model explainability is crucial in scientific applications, and the open sharing of trained models, metrics, and data is essential for driving scientific progress. ML promises to uncover insights in complex biological systems, allowing the automated discovery of novel cellular dynamics.

References:

1. Anjum S., Gurari D. Cell tracking with mitosis detection dataset challenge // Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2020), June 14-19, Seattle, WA, 2020. P. 105-110.

2. Computer Vision and Pattern Recognition in Cornell University // open-access archive. URL: https://arxiv.org/abs/2304.02643 (visiting date: 28.02.2024).

3. O. Gryshkov, M.Tymkovych, O. Avrunin, B. Glasmacher et al. Ethylene glycol improves cryopreservation of cell-seeded electrospun scaffolds in cryobags // Abstracts of 8th European Medical and Biological Engineering Conference (EMBEC 2020), 29 November – 3 December, Portorož, Slovenia, 2020. P. 169.

4. Tymkovych, M., Avrunin, O., Gryshkov, O., Semenets, V., Glasmacher, B. Ice crystals microscopic images segmentation based on active contours . // IEEE 2019. 39-th International Conference on Electronics and Nanotechnology, 16-18 April, Kyiv, Ukraine, 2019. P. 493–496.