A framework for evaluating predicted sperm trajectories in crowded microscopy videos

*David Hart[1], Kylie Cashwell[2], Anita Bhandari[1], *Cameron Schmidt[2],

1 Computer Science Department, East Carolina University, Greenville, NC, USA
2 Biology Department, East Carolina University, Greenville, NC, USA

*Shared senior authorship: hartda23@ecu.edu, schmidtc18@ecu.edu

Abstract

For many years, sperm trajectory analysis for microscopy videos has determined categories of sperm motility. These categories are dependent on kinematic parameters such as Curvilinear Velocity and Beat Frequency. These outputs, however, ultimately rely on the quality of the tracking for each individual sperm in the microscopy video. Previous approaches have relied on calculating these values on short time scales or for isolated sperm to avoid the tracking errors that occur in crowded (high cell-density) microscopy videos, but improving fertility analysis will require accurately tracking trajectories at physiologically-relevant time scales and cell densities. In this work, we identify a framework for accurately assessing the quality of sperm trajectory tracking independent of motility outputs. We utilize common cell tracking metrics and propose modifications based on conditions unique to sperm microscopy. Additionally, we provide a dataset of three fully-labeled microscopy videos to allow for future comparisons and developments.

Author summary

This report develops a computer vision framework to track individual sperm cells in crowded (high cell-density) microscopy videos, with the goal of improving automated analysis of sperm motility patterns. The task of accurately analyzing sperm movement is deceptively challenging due to the high rate of cell crossovers, an issue that has long impeded long-term tracking of sperm trajectories. Here, we introduce new evaluation metrics and provide a labeled dataset to serve as a baseline for future improvements.

Introduction

Inherent difficulty arises when analyzing sperm microscopy videos due to the high density of sperm in a given area. Many works have presented analysis techniques for sperm [1, 2], but these works have focused on analysis of isolated sperm. Standard kinematic parameters and outputs exist [3], but to analyze crowded microscopy videos, many tools require tracking only over short-term paths and aggregating statistics. In comparison, being able to track properties of individual sperm within crowded videos would greatly enhance the understanding of individual sperm motility properties. Such capabilities could have important clinical applications, such as improving selection techniques for In Vitro Fertilization. Ultimately, whether computing standard kinematic

parameters or developing new analyzes, the outputted results will rely on the quality of the tracked sperm trajectories.

In order to effectively track individual sperm in crowded microscopy videos, robust tracking techniques will need to be developed, potentially aided by the onset of AI methods. However, there is currently no established baseline for determining the effectiveness of sperm tracking results. Common metrics and analyses have been established in general cell tracking literature [4–7], but there are unique considerations that need to be addressed for sperm microscopy videos. These metrics also need to account for ambiguities in sperm segmentation, possible crossover of sperm from the 2D perspective, and other unique difficulties that are not common in other cell tracking problems. Additionally, there is little fully-labeled data for sperm videos, requiring *a priori* considerations for evaluating different tracking approaches.

In this work, we establish a framework for evaluating tracking algorithms for sperm microscopy videos. Common cell tracking metrics are adapted to work within the context of sperm microscopy. Given predicted and ground truth trajectories (represented in a simple .csv format), our code computes these metrics to evaluate the performance of the tracker independently of derived kinematic outputs. Establishing this framework will allow for continued development of computer vision approaches that can solve this important task.

As a baseline, we provide three fully-labeled crowded microscopy videos and run the common Trackpy library on them. We report the associated metrics for others to improve upon. We also conduct experiments with an alternative tracker to showcase the effects of poor tracking predictions on metrics and other outputs.

In summary, this work makes the following contributions:

- A framework and relevant metrics for evaluating tracking methods on both labeled and unlabeled data.
- An analysis of tracking quality's effects on the distributions of curvilinear, straight-path, and average velocity.
- A baseline method and a small fully-labeled dataset of microscopy videos to allow for future comparisons.

Additionally, our code will be publicly available upon publication.

Materials and Methods

Participants

This study was approved by the East Carolina University Institutional Review Board and all subjects provided written informed consent prior to participating in research activities. Males ages 18-40 were included in the study. Donors with self-reported sexually transmitted infections within the previous six months were excluded. Baseline fertility was assessed via pre-screening health questionnaire and qualitative semen analysis following World Health Organization Guidelines [8].

Chemicals and Reagents

All chemicals and reagents used in this study were obtained from Sigma Aldrich (St. Louis, MO).

Human Sperm Isolation

Semen samples underwent liquefaction at 37 °C for 30 minutes in a 5% CO_2 incubator. Sperm were isolated from seminal plasma by differential centrifugation in 50% isotonic percoll. Sperm were then washed by centrifugation in modified Biggers, Whitten, and Whittingham media with 3.5% bovine serum albumin. Cells were resuspended in BWW medium for analysis. Cell counts were performed using a hemocytometer. Samples were stored in a 37 °C incubator until imaging. All imaging was performed within 90 minutes of sample collection.

Phase Contrast Video Microscopy

Samples were diluted in BWW medium for imaging. Phase contrast microscopy was performed at 37 °C on a Zeiss Axervert.A1 compound light microscope fitted with a thermal stage (Carl Zeiss AG, Jena, Germany). Cell suspensions were diluted to approximately 8 million cells per mL. Diluted suspensions were added to a 20μm glass depth chamber slide (Hamilton Thorne, Beverly, MA). Phase contrast microscopy videos were obtained at 5X (0.25NA) objective or at 10X (0.25NA). Videos were recorded continuously for two minutes at nine frames per second using Zeiss Zen Lite acquisition software with accompanying time-series data collection module. Raw video files were stored as uncompressed .avi format. Videos for analysis were compressed using h.264 encoding.

Results

Tracking Effects on Standard Sperm Analysis Outputs

In sperm microscopy videos, there are consistent kinematic parameters that are analyzed. These outputs come from the Computer-Aided Sperm Analysis (CASA) system, which defines certain sperm motility and movement patterns [3]. These parameters are:

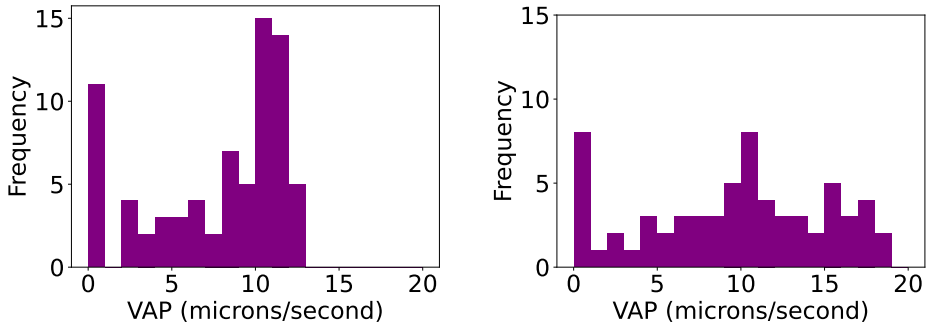
- Curvilinear Velocity (VCL): Measures the total distance traveled along the sperm's actual curvilinear trajectory, expressed in microns per second.
- Straight-Line Velocity (VSL): Calculates the shortest distance between the starting and ending points of the trajectory, divided by the total time, also expressed in microns per second.
- Average Path Velocity (VAP): Represents the average velocity along a smoothed path of the sperm's trajectory, expressed in microns per second.
- Beat-Cross Frequency (BCF): Indicates the frequency at which the sperm's head crosses the average path, measured as a discrete count per second.
- Amplitude of Lateral Head Displacement (ALH): Describes the extent of side-to-side movement of the sperm head along the trajectory. For ALH, the output includes two columns in the .csv file: the maximum and mean amplitudes, expressed in microns.

All parameters are calculated based on the x and y coordinates of the sperm's position in each image frame.

Sperm velocity distributions can reveal important insights into sperm movement patterns and functionality. For example, a bimodal distribution in a 5x magnification suggests the presence of distinct subpopulations, such as progressive, non-progressive, and hyperactivated sperm, while an exponential distribution at a 10x magnification can

indicate that most sperm move slowly, with only a few reaching high velocities. Factors like sperm concentration, medium viscosity, and magnification can shift these various distributions and affect the proportions of motile sperm. Analyzing and understanding these distributions helps explain how sperm function collectively and how selection pressures may favor certain motility traits. By analyzing velocity patterns, it can be determined if sperm are efficiently navigating their environment, indicating the distribution of the population is optimized for fertilization.

The calculations of these distributions, however, ultimately relies on the reliability of the underlying tracking software. In particular, velocity distributions from overcrowded sperm populations can reveal a limit on high-velocity sperm due to trajectory collisions that lead to errors in sperm ID assignment. If there are differences between the actual sperm trajectories and those generated by the tracking algorithms, the presented distributions could drastically misrepresent the underlying behaviors of the sperm population. An example of this phenomenon is illustrated in Figure 1. Thus, to continue to progress sperm analysis techniques, it becomes imperative to evaluate the quality of a tracking algorithm independently of its derived kinematic parameters.



Prediction Actual

Fig 1. Velocity Distribution Comparison Example distributions for the Average Path Velocity (VAP). If a particular tracking algorithm is unsuited to capture high-velocity or crowded trajectories (left), it will misrepresent the motility properties of the sperm population (right).

Evaluating Sperm Detection and Tracking

A microscopy video is made up of individual image frames. The task of effectively tracking sperm throughout the entirety of the video can be separated into two distinct subtasks. First, the sperm must be detected in each individual frame. Second, individual sperm must be correctly identified and linked to detected sperm in subsequent frames. This process has many similarities to general cell tracking. Thus, many of the metrics used to evaluate the effectiveness of cell tracking can be used in sperm tracking.

Of particular interest are metrics from two common benchmarks: The Mutli-object Tracking Benchmark [5] and the Cell Tracking Challenge [4]. General multi-object tracking has context outside of the biological realm and well-established metrics have been used such as MOTA [9], IDF1 [10], and HOTA [11]. These general benchmarks are calculated based on both the accuracy of the detections (true positives, false positives, and false negatives) and the number of correctly and incorrectly linked objects between frames. For cell tracking, metrics are often defined based on the spatio-temporal graph representation of the trajectories as first outlined in [12]. Each metric can be thought of as the number of steps needed to reconstruct the ground truth graph given the predicted graph. DET and TRA are respectively based on the detection and overall

tracking pieces of the reconstructions. Another metric of interest is track fractions (TF) which describes the average length of trajectory that is correctly identified by a predicted trajectory. Additional details providing the mathematical definitions of each metric are provided in Appendix S1.

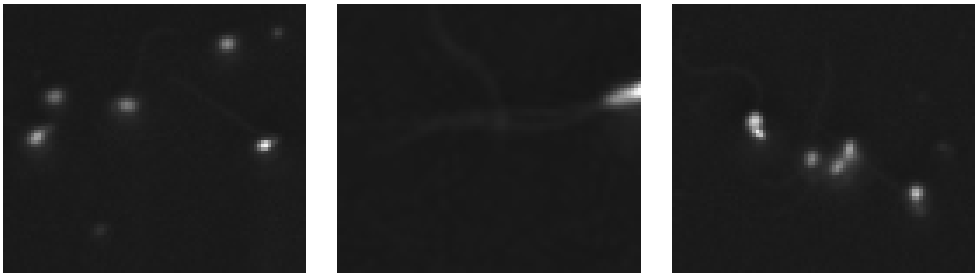
These common cell tracking metrics provide a foundation for evaluating quality of the sperm tracking algorithms. Sperm tracking, however, has unique differences from standard cell tracking that must be considered and properly accounted for. We will describe specific considerations for each subtask in the following subsections.

Unique Considerations for Evaluating Sperm Detection

Unlike most cells, sperm are incredibly thin, especially along the flagellar tail. Even at high magnifications and resolutions, sperm are usually a few pixels wide, but many pixels long. In most cell-tracking detection metrics, it is assumed that a segmentation or a bounding box is provided during the detection process. This is needed to accurately compute the Jaccard similarity scores (Intersection over Union) to match predicted trajectory IDs to the reference IDs [4]. Relying on these similarity scores for sperm trajectories would provide inconsistent association scores because of the ambiguity in labeling sperm pixels. Additionally, there are multiple confounding factors that can throw off the detection process when compared to many cell tracking setups:

- Variations in brightness of pixels within a video can make some sperm more visible than others and make sperm outlines unclear.
- Dust, dead cells, and other debris in the video may appear like sperm, but ultimately do not move within the video.
- Sperm may enter and exit the frame, and sperm near boundaries of the video may be partially missing in shape.
- Sperm can overlap each other as they swim above or below each other, making them appear as one or more sperm of irregular shape.
- Labeling programs may be inconsistent in the use of points, bounding boxes, or segmentations for identifying sperm, as well as differences in the labelers isolating the full body or the sperm head only.

Examples of these various issues are visualized in Fig. 2



Brightness Variations | Border Cutoff | Overlap

Fig 2. Sperm Detection Issues Example issues with detection of sperm. Variations in brightness and dead cells in view (left), partially missing sperm near boundaries (middle), overlap and irregular shapes (right).

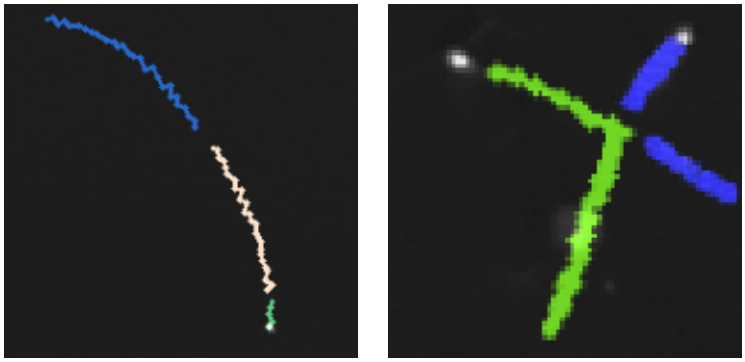
To combat these issues and establish correct predicted-to-actual ID linking for all types of sperm trajectory videos, it is reasonable to represent sperm detection using a

single point rather than a segmentation. The x,y coordinate is approximately centered on the head of the sperm which we call the sperm *centroid* (since most of the mass is located in the sperm head). Localization between the actual and predicted point can be quickly computed using the Hungarian algorithm with a distance cutoff. The standard cell tracking metrics can be then used in the same fashion as before.

Unique Consideration for Evaluating Sperm Tracking

Once sperm have been identified in all frames of the video, they must be properly linked together to make a complete tracking of individual sperm. Some approaches may also rely on detections in previous frames to update detections in future frames and link in a forward fashion. Regardless of the approach, the goal is assign each sperm with a unique identifier that it keeps throughout the whole microscopy video. Doing so enables drawing the trajectory path of each sperm and calculating the kinematic parameters.

Tracking sperm presents unique challenges compared to most cell tracking problems. First, many sperm move very quickly compared to spatial resolution of most microscopy videos. Even during high FPS recordings, some sperm may move multiple pixels per frame. For tracking algorithms, this can cause frequent dropout and relabeling of sperm in subsequent frames. Second, as mentioned previously, sperm paths can cross over each other as one sperm swims above another. This crossover can cause an incorrect path labeling, where the two sperm swap trajectory labels in subsequent frames. These two issues are illustrated in Figure 3.

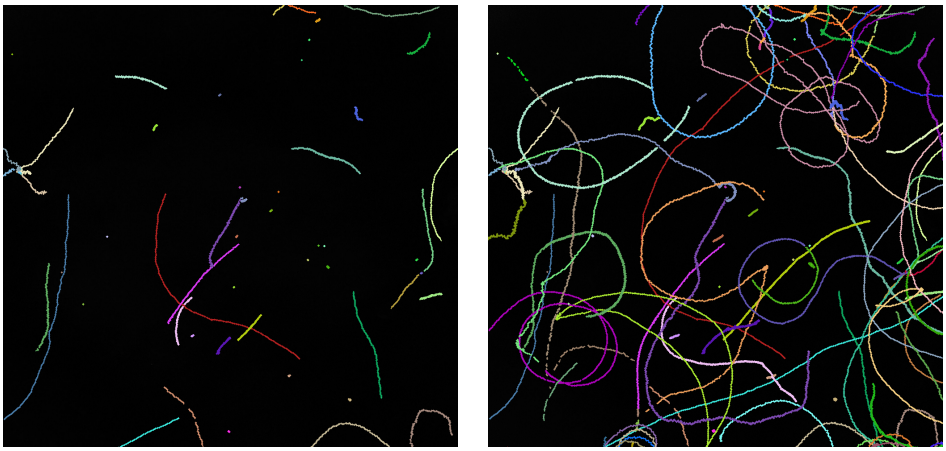


Dropout Crossover

Fig 3. Sperm Tracking Issues Example visualization of sperm paths issues. A single sperm may appear to dropout for a few frames, causing the algorithm to think that new sperm are appearing in and out of view (left). Two sperm that intersect may cause the path of one sperm to be assigned to another, causing a straightline path to be incorrectly marked as jagged (right).

In addition to these challenges, there are additional considerations given sperm analysis use-cases. First, in many instances, nonmoving or dead sperm are not of interest to the analysis. Nonmoving sperm are easy to identify in each subsequent frame and may disproportionally effect the tracking metrics, making a tracking algorithm appear to perform better than it actually is. Additionally, some nonmoving dust and debris may look like sperm, and be incorrectly labeled on reference datasets. Second, capturing long-term movements are usually much more important for understanding macroscopic patterns and behaviors of individual sperm. An illustration comparing the short-term and long-term trajectories is shown in Figure 4.

An effective tracking algorithm would be able to account for these confounding factors, as well as favor algorithms that capture faster moving objects and longer paths.



Short-term Paths Long-term Paths

Fig 4. Sperm Tracking Example Example visualization of sperm paths for a single video. In the short term, sperm paths do not present much information (left). It requires long-term tracking to determine unique motility properties (right).

Thus, when labeled ground truth data is present, the standard tracking metrics are used. However, we propose an additional filtered version of these metrics which removes non-moving and very-short paths. For these metrics, a preprocessing step is taken for both the ground truth and predicted paths. This preprocessing step removes all paths that do not spatially extend pass a small radius of ϵ. This removes all paths for nonmoving or nearly nonmoving sperm, giving more weight on the metric to faster moving sperm and longer-term trajectories.

Comparing Algorithms on Unlabeled Data

The most effective way to evaluate detection and tracking algorithms is to compare predicted outputs against labeled ground truth data. However, labeling sperm microscopy videos, especially in highly crowded scenarios, is difficult and time-consuming. It is worth considering whether indirect comparisons for evaluating methods is possible in the absence of labeled data. We investigate two possible metrics to compare the performance between tracking algorithms without labeled data.

Sperm Per Frame - While some sperm enter and exit the field of view in each frame, the number of sperm should stay relatively consistent throughout the video as the number of sperm that enter and exit the frame should average out. The number of detections in each frame generates a distribution of values, whose mean should approximate the number of visible sperm in each frame of the video. The standard deviation should tend towards zero as the accuracy of the detection improves. An example visualization of the sperm per frame distribution is shown in Figure 5.

Frame Per Sperm - Even for microscopy videos with shorter lengths, many sperm do not stay within view for the entire video. Fast moving sperm tend to exit after a short time. The number of frames that each sperm ID exists generates a distribution of values as shown in Figure 5. As can be seen, this distribution does not necessarily approximate a normal distribution, making it unreasonable to make assumptions about the expected mean for any video. However, most trackers have a tendency to dropout sperm for frames in the video, rather than track for too long. Thus, a well-tracked video should tend to have sperm IDs that exist for a longer number of frames in the video.

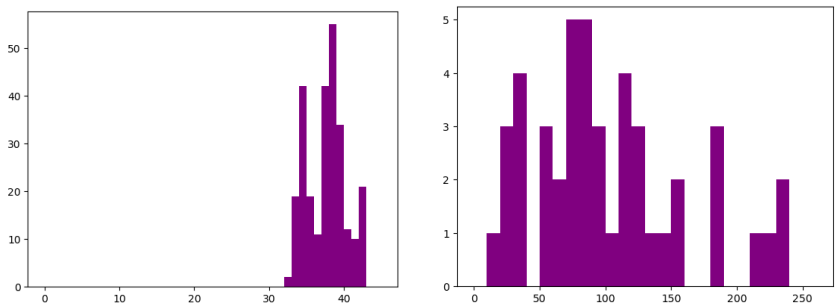


Fig 5. Sperm and Frame Distributions Visualizations of the Sperm Per Frame and Frames Per Sperm distributions for the filtered ground truth data in Video 3. The Sperm Per Frame tend to approximate a normal distribution. The Frames Per Sperm has a peak at a lower value (most fast moving sperm exit the frame) with values that go up to the full length of the video.

Evaluation on the Dataset

With the evaluation process established in the previous section, the detection and tracking metrics are presented for a baseline tracking method on a dataset of three microscopy videos that were collected from human sperm samples. For a baseline detector, we use the particle detector *locate* function from the Trackpy library [13]. For all experiments, we use default parameters of an 11-pixel diameter and 500 minmass. For a baseline tracker, we use particle tracker *predict* function from the Trackpy library [13]. For all experiments, we use default parameters of 25 for the search range and 3 for the frame memory.

In the following subsections, we present the labeling process for the dataset, outputted metrics for the baseline Trackpy method, and experiments on how the metrics can indicate the reliability of standard motility outputs.

Labeling a Microscopy Dataset

Sperm samples were collected from human participants. The samples were washed and diluted before recording the microscopy videos at a 10x magnification with a negative phase contrast. Additional details are outlined in the Materials and Methods section.

To create the fully-labeled sperm trajectory data, the sperm microscopy videos were first fed through the baseline Trackpy approach. The resulting tracks ultimately had errors that needed to be corrected. To correct the mistakes, a tool was built using OpenCV [14] in Python that allows a human labeler to manually modify tracks. Simple mouse and keyboard commands for adding, deleting, and merging trajectory data were included. The labeler program is included as part of the code repository.

Metric Computations

For our human-labeled dataset, we use the baseline Trackpy approach and report the common cell tracking metrics on the three videos. The ID association process described previously is used. Additionally, we compare results with and without filtering out the non-moving objects in the calculations. The results are given in Table 1. For all metrics, a higher score indicates better performance and 1.0 is the highest score.

The outputted metrics serve as a baseline for future tracking algorithms to compare against. As more tracking algorithms are compared and their strengths analyzed,

additional studies can be conducted to determine which metrics are most useful for the sperm tracking problem.

Table 1. Detection and Tracking Results

	Video 1		Video 2		Video 3	
Metrics	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered
DET	0.970	0.960	0.979	0.976	0.967	0.924
TRA	0.796	0.791	0.805	0.803	0.791	0.760
TF	0.891	0.881	0.923	0.920	0.841	0.804
MOTA	0.789	0.737	0.822	0.790	0.840	0.818
IDF1	0.816	0.782	0.867	0.846	0.839	0.795
HOTA	0.831	0.801	0.878	0.860	0.856	0.819

Table notes: Common multi-object and cell-tracking metrics computed for the three videos. For all metrics, a higher score indicates better performance and 1.0 is the highest score. The baseline Trackpy approach is measured against the human-labeled and corrected data. Additionally, the results are also compared before and after filtering the predictions and labels to exclude nonmoving objects, better capturing the trackers ability to follow fast moving objects.

Relating Tracking to Motility Outputs

To demonstrate the influence of the tracking method on the motility outputs, we intentionally make a second algorithm that under performs on the tracking task. This alternate approach uses the same Trackpy detect and tracking methods, but uses a search range of 7 pixels instead of 25. This greatly limits the algorithms ability to track fast-moving sperm. As we will show, many inaccuracies in the tracking can occur from just one unoptimized hyperparameter.

First, we present the detection and tracking metrics, comparing the alternate approach to the baseline approach after filtering. These results are presented in Table 2. As is shown, the alternate approach underperforms in every metric, indicating that the metrics adequately represent fast-moving objects.

Table 2. Detection and Tracking Ablation

	Video 1		Video 2		Video 3	
Metrics	Alternate	Baseline	Alternate	Baseline	Alternate	Baseline
DET	0.624	0.960	0.634	0.976	0.713	0.924
TRA	0.511	0.791	0.524	0.803	0.583	0.760
TF	0.514	0.881	0.588	0.920	0.638	0.804
MOTA	0.509	0.737	0.444	0.790	0.633	0.818
IDF1	0.564	0.782	0.648	0.846	0.669	0.795
HOTA	0.618	0.801	0.686	0.860	0.707	0.819

Table notes: Comparison of tracking and detection metrics between the alternate approach and the baseline on the filtered data. The alternate approach is deliberately implemented to fail in tracking fast-moving objects, showcasing the metrics effectiveness in the sperm tracking task.

Second, we present the Sperm Per Frame and Frames Per Sperm metrics on the alternate approach, baseline approach, and ground truth labels for the three videos on the filtered data. These measurements are presented in Tables 3 and 4 respectively. The

results indicate that Sperm Per Frame may not be a fair comparison metric between algorithms. The baseline approach has values that are higher than the ground truth trajectories, meaning the baseline approach likely has spurious detections. Since the ground truth data has values between the baseline and alternate approaches, it is likely unreasonable to correlate sperm per frame with tracking performance in general. In comparison, the Frames Per Sperm appears to be reasonably correlated with performance. Higher frames per sperm could generally indicate a better tracking algorithm, even in the lack of labeled data. Though, as is shown for Video 1, some exceptions do occur.

Table 3. Sperm Per Frame Metric

Sperm Per Frame	Video 1	Video 2	Video 3
Alternate Approach	25.27	18.56	21.50
Baseline	40.94	25.82	38.96
Ground Truth	33.32	21.50	37.16

Table notes: Average Sperm Per Frame for each method on each video using the filtered trajectory data. Note that the ground truth results are between the poorly performing alternate approach and the baseline approach, indicating that Sperm Per Frame is not a reliable measure of performance.

Table 4. Frames Per Sperm Metric

Frames Per Sperm	Video 1	Video 2	Video 3
Alternate Approach	71.72	128.94	106.61
Baseline	144.11	154.91	138.71
Ground Truth	137.45	157.63	152.65

Table notes: Average Frames Per Sperm for each method on each video using the filtered trajectory data. The higher the Frames Per Sperm, the more likely the better performance of a method, but as illustrated in Video 1, this is not always the case.

Third, we present the computed motility outputs for the alternate approach, baseline, and ground truth. These distributions are shown in Figure 6. As can be seen, these motility outputs are drastically effected by the tracking approach. Specifically, the alternate approach can not track the trajectories of fast moving sperm, greatly skewing the resulting distributions and affecting the predicted motility categories. Overall, this indicates the need for open tracking metrics to evaluate the reliability of these values.

Discussion

In this paper, we have established a framework for the evaluation of detection and tracking algorithms for sperm microscopy videos. As computer vision, machine learning, and artificial intelligence techniques continue to improve tracking algorithms, this evaluation framework can serve as a basis for fair comparison.

With that goal in mind, we aim to make this code publicly available for other to use and run their methods on. Our code will be available upon publication. For a given video, trajectory data can be stored in a single comma-separated-values file (.csv). This .csv file contains x,y coordinates, a sperm label, and frame in each row. This provides an efficient, human-readable, and unambiguous way of storing predictions or ground truth data. Once the prediction data is provided in a standard format, the labeled and

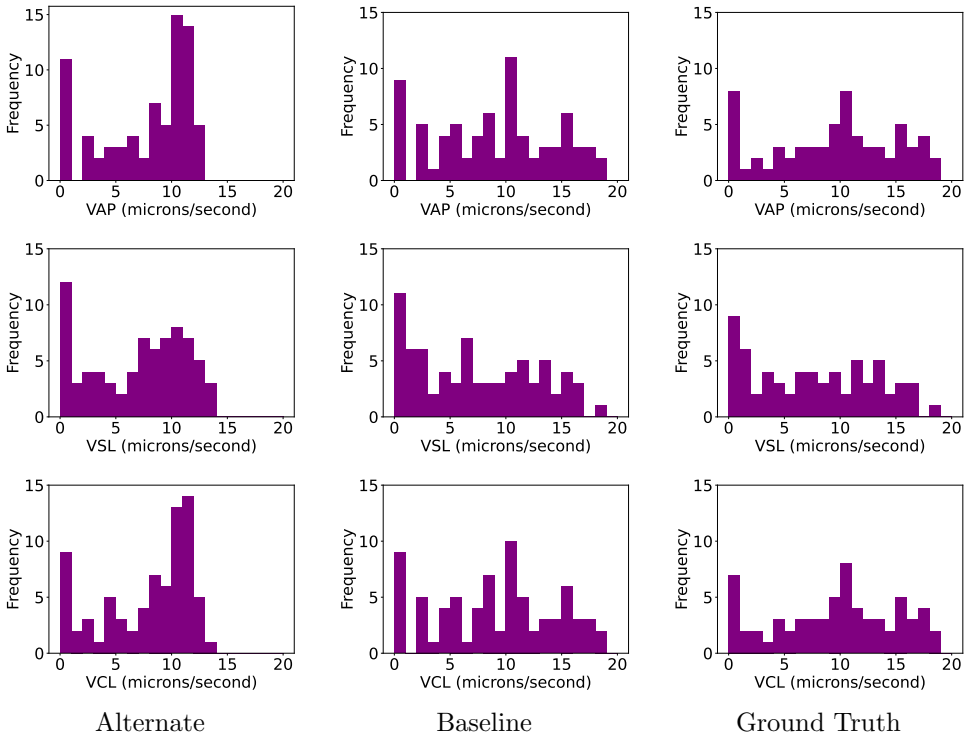


Fig 6. Motility Outputs The outputs for VAP, VSL, and VCL on the alternate approach, baseline tracking approach, and ground truth for the filtered data of Video 3. Notice how the alternate approach fails to capture the fast moving velocities, which would greatly alter the predicted motility categories.

unlabeled metrics can be computed to evaluate its performance as previously described. If no ground truth is provided, only the unlabeled metrics are computed.

Conclusion

In this work, we have presented a framework for effectively evaluating tracking algorithms for sperm microscopy videos. We have considered the challenges and difficulties of this task and proposed modifications to standard metrics to best cater to the unique circumstances of sperm microscopy. We have established a baseline approach and dataset for others to compare tracking algorithms against. This framework will allow for the continued development of effective sperm analysis tools that can utilize new computer vision and artificial intelligence techniques.

In future work, we aim to expand our comparisons to additional datasets such as VISEM-Track [15] and MIaMIA-SVDS [16]. Each dataset stores the trajectory information in a different way, so tools will need to be develop to convert between representations. Additionally, though the segmentation is not ultimately needed for trajectory tracking, obtaining full sperm outlines can be helpful for biological outputs such as flagellar waveform analysis. Additional baselines and metrics can be established to properly account for these use cases.

Supporting information

S1 Appendix. Common Cell Tracking Metrics

For convenience of the reader, we provide a description of the common multi-object and cell tracking metrics. Relevant works are cited to provided sources for additional details.

Association Accuracy and Target Effectiveness

Association Accuracy (AA) and Target Effectiveness (TE) have been a standard metric within the cell tracking literature. AA is a measure of the number of correctly predicted cell associations between frames divided by the total number of associations in the ground truth data. TE is a measure describing the number of frames a ground truth track is correctly followed by the predicted tracks. These two metrics are still used frequently to compare modern cell tracking approaches [17–19].

MOTA, IDF1, HOTA, and CHOTA

Multi-object tracking has many applications outside the realm of cell tracking. Standard metrics have been used for comparing methods on various tasks, such as in the Multi-object Tracking Benchmark [5]. These metrics have also been adopted as cell tracking for comparison, such as in the Cell Tracking With Mitosis Detection Dataset Challenge (CMTC) [20].

MOTA is a measure of the total detection coverage, and is calculated as

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t} \tag{1}$$

where t is given frame of the video, GT is the number of objects in the ground truth of the frame, FN is the number of false negatives (missed detections), FP is the number of false positives (spurious detections), and $IDSW$ is number of correct detections that are labeled with a different ID than the track they were associated with in the previous frame. The best MOTA score is 1.0 and scores can go into the negatives.

IDF1 is a measure of the consistency of labels across frames of a video. It uses the Hungarian algorithm to optimize how long each ground truth track was connected to a single ID trajectory. After this matching, the metric is calculated as

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{2}$$

where IDTP is the total length and number of correctly labeled tracks for all frames, IDFP is the number of predicted tracks that don't match any ground truth, and IDFN is and ground truth tracks that weren't predicted.

Recently, a new metric has been proposed called HOTA that better balances detection and association pieces [11]. A variant of this metric has been proposed specifically for cell tracking called CHOTA [21], but CHOTA was designed specifically to measure mitosis splitting events which do not occur in sperm.

DET, TRA, and TF

The Cell Tracking Challenge [4] has adopted standard evaluation methodology that can be used across many types of cells and data types. Many of these metrics rely on the Acyclic Oriented Graph Matching (AOGM) measure that is presented in [12]. This measure determines the number of steps needed to correct a graph representation of the predicted tracks to match the graph representation of the ground truth tracks. It is calculated as

$$AOGM = w_{NS}NS + w_{FN}FN + w_{FP}FP + w_{ED}ED + w_{EA}EA + w_{EC} \quad (3)$$

In this formulation, FN and FP represent false negatives and false positive in detections. ED, EA, and EC represent different kinds of incorrect tracking associations (edges in the graph). w represents a weight applied to each measure based on how difficult it would be to correct an error through a manual modification. The standard values are $w_{NS} = 5, w_{FN} = 10, w_{FP} = 1, w_{ED} = 1, w_{EA} = 1.5, w_{EC} = 1$. NS represents mitosis events where cell splits, so this term can be ignored for sperm tracking.

The DET and TRA metrics derive from this measure. As an example, TRA is calculated as

$$TRA = 1 - min(AOGM, AOGM_0)/AOGM_0 \quad (4)$$

where $AOGM$ is the value of the metric for the predicted tracks, and $AOGM_0$ is the value of the metric for an empty reference (i.e. creating all tracks from scratch). This forces the measure to be between 0-1. DET is calculated in the same way but with no weight assigned to the edge calculations ($w_{ED} = w_{EA} = w_{EA} = 0$).

The Cell Tracking Challenge [4] also presents other biological plausibility metrics, many related to cell mitosis. Track Fractions (TF) is of particular interest for the sperm tracking problem. For each ground truth track, the predicted track that matches for the most frames is determined. Then, the number of frames of that are correctly matched is divided by the total frames for the ground truth track. These are then averaged for all ground truth sperm trajectories.

Jaccard Similarity/IoU/SEG

Whether as a bounding box or a pixel segmentation, the accuracy of the overlap between predicted and ground-truth regions is usually calculated using Jaccard Similarity. This is also known as Intersection over Union (IoU). This is calculated as

$$IoU = \frac{R \cap S}{R \cup S} \quad (5)$$

where R is the ground truth segmentation and S is the predicted segmentation. The SEG measure from the Cell Tracking Challenge [4] is also based-on this score. In this work, we do not provide segmentation results do due the difficulty of manually labeling each pixel. Thus, this metric is not presented in this work.

References

1. Hansen JN, Rassmann S, Jikeli JF, Wachten D. SpermQ–A Simple Analysis Software to Comprehensively Study Flagellar Beating and Sperm Steering. Cells. 2019;8(1). doi:10.3390/cells8010010.

2. Fujii T, Nakagawa H, Takeshima T, Yumura Y, Hamagami T. Automated Sperm Assessment Framework and Neural Network Specialized for Sperm Video Recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2024. p. 7675–7684.

3. Alquézar-Baeta C, Gimeno-Martos S, Miguel-Jiménez S, Santolaria P, Yániz J, Palacín I, et al. OpenCASA: A new open-source and scalable tool for sperm quality analysis. PLoS computational biology. 2019;15(1):e1006691.

4. Maška M, Ulman V, Delgado-Rodriguez P, Gómez de Mariscal E, Necasova T, Guerrero Peña FA, et al. The Cell Tracking Challenge: 10 years of objective benchmarking. Nature Methods. 2023;20:1–11. doi:10.1038/s41592-023-01879-y.

5. Dendorfer P, Osep A, Milan A, Schindler K, Cremers D, Reid I, et al. MOTChallenge: A Benchmark for Single-camera Multiple Target Tracking. International Journal of Computer Vision. 2020; p. 1–37.

6. Bragantini J, Lange M, Royer L. Large-Scale Multi-Hypotheses Cell Tracking Using Ultrametric Contours Maps. 2023;.

7. Bragantini J, Theodoro I, Zhao X, Huijben TA, Hirata-Miyasaki E, VijayKumar S, et al. Ultrack: pushing the limits of cell tracking across biological scales. bioRxiv. 2024; p. 2024–09.

8. Organization WH, et al. WHO laboratory manual for the examination and processing of human semen. World Health Organization; 2021.

9. Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing. 2008;2008:1–10.

10. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. Springer; 2016. p. 17–35.

11. Luiten J, Osep A, Dendorfer P, Torr P, Geiger A, Leal-Taixé L, et al. Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision. 2021;129:548–578.

12. Matula P, Maška M, Sorokin DV, Matula P, Ortiz-de Solórzano C, Kozubek M. Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. PloS one. 2015;10(12):e0144959.

13. Allan DB, Caswell T, Keim NC, van der Wel CM, Verweij RW. soft-matter/trackpy: v0.6.4; 2024. Available from: `https://doi.org/10.5281/zenodo.12708864`.

14. Bradski G. The OpenCV Library. Dr Dobb's Journal of Software Tools. 2000;.

15. Thambawita V, Hicks SA, Storås AM, Nguyen T, Andersen JM, Witczak O, et al. VISEM-Tracking, a human spermatozoa tracking dataset. Scientific Data. 2023;10(1):260.

16. Group M. MIaMIA-SVDS. 2021;doi:10.6084/m9.figshare.15074253.v1.

17. Hayashida J, Nishimura T. MPM: Joint Representation of Motion and Position Map for Cell Tracking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020;.

18. Nishimura K, Hayashida J, Wang C, Ker DFE, Bise R. Weakly-Supervised Cell Tracking via Backward-and-Forward Propagation. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII. Berlin, Heidelberg: Springer-Verlag; 2020. p. 104–121. Available from: `https://doi.org/10.1007/978-3-030-58610-2_7`.

19. Ben-Haim T, Riklin-Raviv T. Graph Neural Network for Cell Tracking in Microscopy Videos. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022.

20. Anjum S, Gurari D. CTMC: Cell Tracking With Mitosis Detection Dataset Challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2020.

21. Kaiser T, Ulman V, Rosenhahn B. CHOTA: A Higher Order Accuracy Metric for Cell Tracking. In: European Conference on Computer Vision Workshops (ECCVW). Springer; 2024.