A Review of Tracking Algorithms Applied to Deep Learning Annotations of Spermatozoa

P. Grant M. Nelhams N. Wray S. Hoeberichts R. Bergna

March 8, 2021

EMAT30005 Mathematical and Data Modelling

Department of Engineering Mathematics, University of Bristol

1 Introduction

There is a growing need for successful male infertility treatments due to measurable trends in decreasing sperm quality, namely 'low concentration, poor motility or abnormal morphology' [1] in recent decades. The decline has been observed globally across populations [2][3][4] and age groups [5]- often attributed to environmental factors. Furthermore, the mean age of fatherhood has increased in the western world [6][7][8] whilst evidence indicates that the same measures of sperm quality, and thus fertility, are inversely proportional to the subject's age [9][10][11].

Male infertility treatments require an extensive analysis of sperm samples, but current practises are limited by technology and have yet to accommodate modern advancements. Traditional treatments suffer from an over-reliance on manual processing and evaluation by technicians and researchers, as outlined by The World Health Organisation [12], which is 'laborious and subjective' [13] and 'requires extensive training' [14]. Consequently, this has led to the rise of computer-aided sperm analysis (CASA), which uses time-lapse microscope image processing to detect, track and classify sperm for kinematic and physiological analysis. Despite the promise of rapid, multi-object scalable analyses, CASA suffers from significant drawbacks due to tracking errors caused by sperm collisions and foreign particles in spermatozoa samples[14]. Since faster sperm are more likely to collide, their tracks are more likely to be excluded from CASA analysis which causes bias in motility measurements [13]. Additionally, to avoid collisions, samples are 'often diluted or analysis is limited to short video clips' [13][12] preventing a rigorous and objective appraisal of the original sample.

This report aims to compare various algorithms' effectiveness of sperm tracking using secondary data where a deep neural network has made sperm head detections. An effective algorithm must form sperm tracks from sequential video frames. More importantly, it must associate disjoint tracks of the same sperm caused by detection inaccuracies, despite a noisy input signal and cell collisions. Whilst particle tracking research is substantial, many biological applications are single target [15] [16] or require fluoresce imaging [17] [18] [19] that is inappropriate to use when selecting sperm for fertility treatment purposes. Recent adaptations of tracking algorithms such as the Joint Probabilistic Data Association Filter [13] and a less computationally costly single frame-differencing method [1] show great promise. However, neither involves reconciliation of deep learning annotations, and the former involves significant pre-processing and thus is not directly comparable to the data of this report.

The data of interest is from samples of two patients: patient 49 with a low sperm concentration and patient 57 with a moderately high sperm concentration. For each patient, two aliquots from one sample were taken and analysed separately into ten, five-second videos (covers). The data is used in this report as follows: Section 2 presents the ground-truth tracks as a baseline and for later reference, in Section 3 a range of common clustering algorithms are employed and evaluated for their performance of sperm track identification and association, Section 4 proposes original extensions to sperm clustering to deal with discontinuities, Section 5 implements reputed multi-target trackers and Section 6 reflects on data quality, limitations of the report and proposes sources of further research.

To compare the performance of the tracking algorithms, the cover sample 04 of patient 49 is used throughout this report to present any findings. This specific dataset was chosen because it has few anomalies such as fluid flow or foreign particles while still having enough motile sperm to test the algorithms' limitations. Various metrics were implemented to analyse the results. The industry-standard metric, known as OSPA [20], or other accuracy measures such as MOTA [21] require knowing the ground-truth; thus, additional metrics were proposed since obtaining the ground-truth for all the sample data is infeasible.[1]

2 Manually tracking sperm using AI-based detections

Having the ground truth allows for more precise insights into the effectiveness of each of the models. Manual analysis is currently the most accurate method, but it is also the slowest. Gathering the ground truth is not scalable; if it were

[1] Each tracking algorithm is presented as three, two-dimensional graphs with clusters uniquely coloured. Key tracks are directly addressed in the figure annotations, rather than by colour, to account for colour-blindness.

possible to hand-cluster points efficiently, then the machine-learning acceleration would be unnecessary. Nonetheless, by hand-clustering a key chaotic sperm sample with assistance of the original annotations (patient 49, cover 04), inferences for why each model succeeds or fails can be gathered and later generalised for all of the samples.

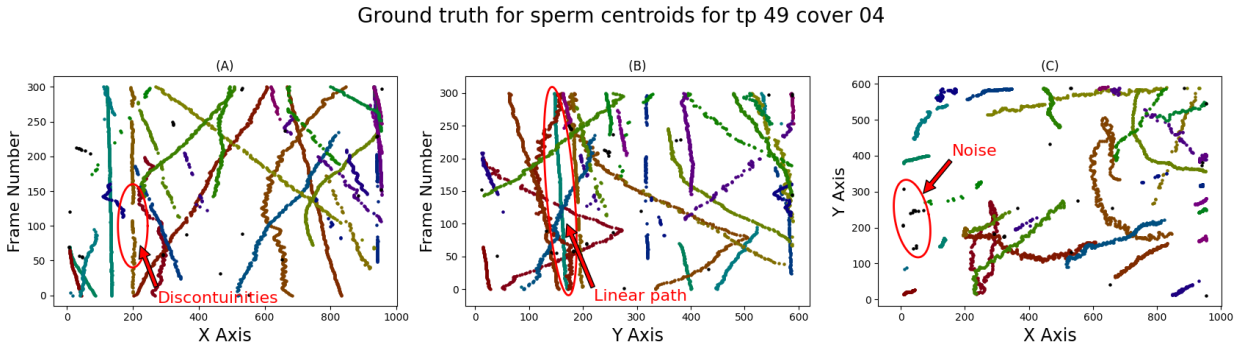


Figure 1: The results for manual tracking are shown above. The debris misclassified as sperm in the videos have been labelled as the final cluster, 32, shown in black in Subfigure C. Each frame was visually examined with the bounding boxes overlaid to provide a starting point for inspection. Subfigures A and B display the 2D projections of the sperm cells over time. Subfigure C highlights the spatial coordinates exclusively. Subfigure A shows an example of discontinuous points of the same trail. Subfigure B displays a linear path, obscured slightly by other sperm trails.

3 Basic clustering applied to AI-based detections

Sperm tracking in three dimensions can be treated as a clustering problem, provided that the sperm head detection is always accurate and consistent. The sperms are expected to move slowly in comparison to the frame rate, which means that methods which effectively cluster linear data may effectively cluster slow, linear samples. However, Figure 2 shows that the sperms quickly deviate from linear paths. The popular basic clustering algorithms considered were density-based spatial clustering (DBSCAN) [22], hierarchical-DBSCAN (HDBSCAN) [23] and Guassian-Mixture Modelling (GMM) [24].

3.1 Simple clustering algorithms

The K-means algorithm ineffectively tracks sperm cells, due to how the algorithm always converges to oblate spheroid cluster shapes [25]. The inaccurate clustering can be seen below in Figure 2. An additional problem is that K-means will cluster every point. Consequently, noise is unaccounted for in most cases and these outliers are included in sperm tracks.

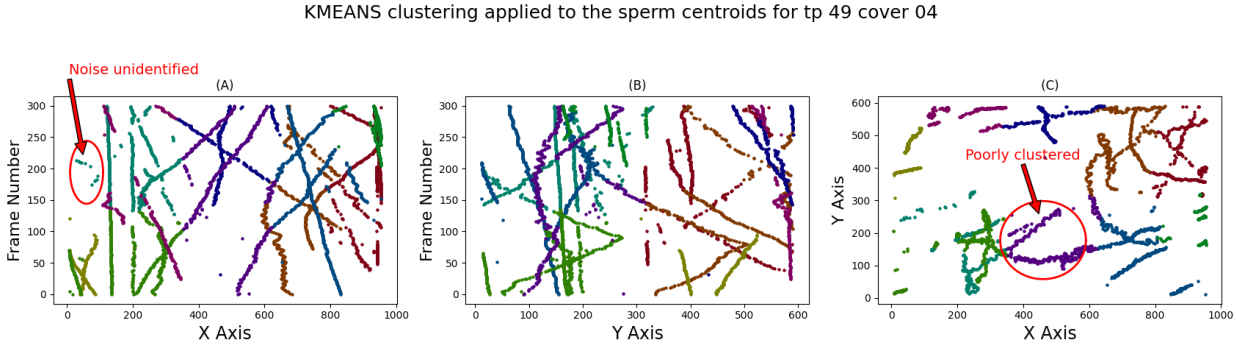


Figure 2: Since the K-means algorithm is unable to identify noise, as shown in Subfigure A, all noise is included within adjacent tracks. Worse still, the K-means algorithm converges to spherical clusters, so any points within the spheres may be misclassified.

Assuming that the sperms are continuously detected from the videos, DBSCAN and HDBSCAN would be effective clustering algorithms for producing accurate tracks. DBSCAN clusters together neighbouring points, within a threshold distance ϵ. The threshold distance should be set to the mean distance a sperm can travel between frames and the number of required neighbours should theoretically be set to three, one each for the previous frame, current frame and the following frame. HDBSCAN does the same for automatically varying threshold ϵ [26]. The assumption breaks down across the samples, because not all sperm cells can be continuously detected with complete certainty across every frame and furthermore sperms may accelerate, generating large discontinuities and misclassifications.

As seen above in Figure 2, there are major discontinuities in the sperm tracks, attributed to losses in detection. Consequently, DBSCAN forcibly chooses between producing broken clusters or combining two or more separate sperm tracks, however both of these outcomes are undesirable. Broken clustering will artificially increase the sperm count, decreasing the average motility away from the ground truth. Combining sperm tracks artificially increases the average motility and it is impossible to classify sperms when the tracks are muddled. For consistent sperm motility analysis and classification it is better to have broken clusters than misclassified clusters.

Results for DBSCAN and HDBSCAN can be seen in Figures 3 and 4. HDBSCAN produces overly many clusters, but it tends not to misclassify clusters: any near-intersections of clusters are bridged by DBSCAN, but not by HDBSCAN. For all of the training data, the number of sperms in the video n_s exists within the range $1 \leq n_s \leq n_{HDBSCAN}$. The last popular clustering algorithm considered was GMM with its results below in Figure 5.

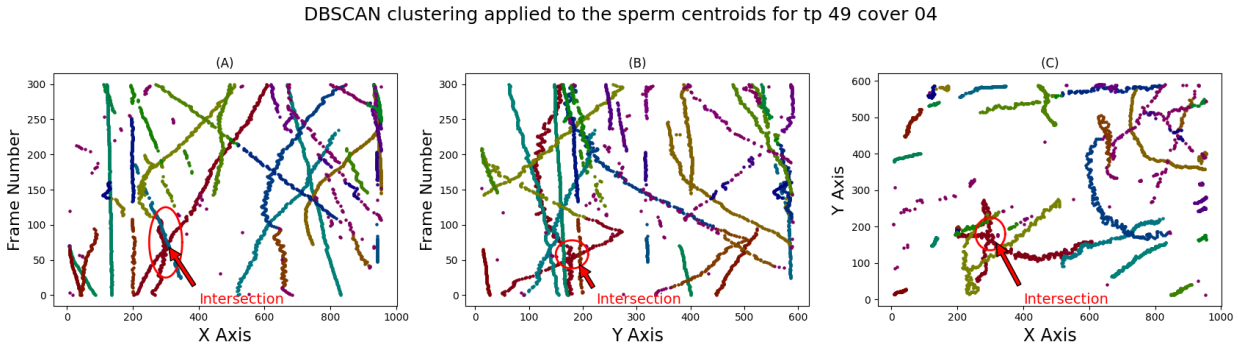


Figure 3: The discontinuities in tracks produce new clusters for the same sperm track in some cases. Worse still, in many cases, such as indicated on the graph, nearby sperm tracks are incorrectly combined due to intersections.

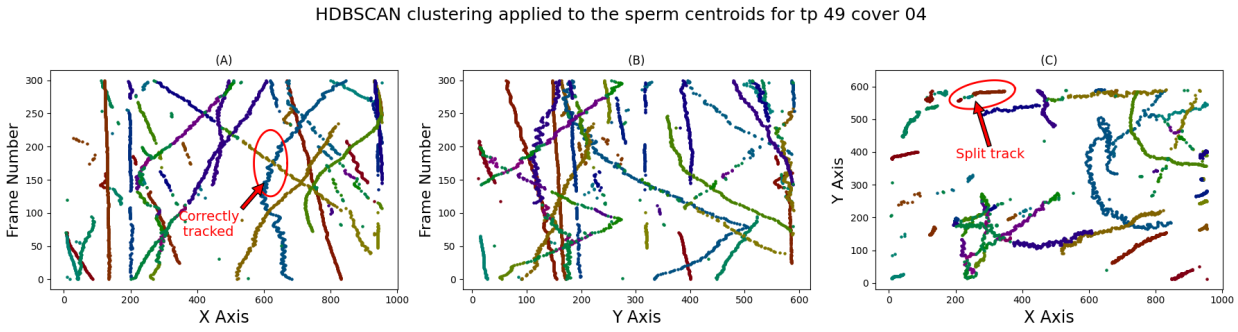


Figure 4: Many sperm intersections are resolved accurately and few sperm tracks are wrongly combined, however, the issue with discontinuities and track splitting is worsened for HDBSCAN as displayed in Subfigure C. There are now 65 clusters compared to the 49 above in Figure 3.

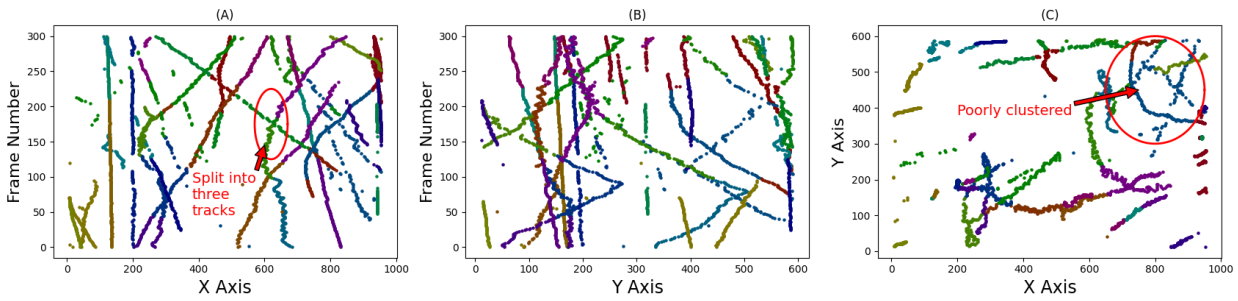


Figure 5: Trialling with changing the number of clusters was done between $1 \leq k \leq n_{\,HDBSCAN}$ and the most optimal clustering formation is shown above for 10 trials for each n. It follows experimentally that optimal results for gaussian mixture modelling are obtained by selecting the number of clusters as around $k = \lceil \frac{n_{\,HDBSCAN}}{2} \rceil$.

Many sperm are misclassified with GMM, comparable to that of K-means in Figure 2. Consequently, GMM and K-means are poor approaches to tracking, worse still they are both stochastic, so their results will vary each time. HDBSCAN is a good approach for determining the upper bound of the problem. DBSCAN produces either many small accurate sub-tracks or few inaccurate combined clusters, both of which are better than the other clustering methods, but this is still insufficient for analysis.

3.2 Stable nearest-neighbour search

None of the aforementioned methods are capable of ensuring that each cluster does not overlap sperm tracks. By adapting the nearest-neighbour search algorithm [27] to map only between every consecutive frame, it is possible to guarantee a stable clustering algorithm that will never overlap sperms. This method shall be referred to as the 'closest frame' clustering technique. The time and space complexity for this algorithm is $O(n)$ for n datapoints, with derivation in the appendix D.1.

Since the algorithm never misclassifies multiple sperm tracks as a singular track, the number of clusters will always exist within the range $1 \leq n \leq n_{closest-frame}$ and the following is true for all the training data, but not necessarily true for all possible data: $n_{HDBSCAN} \leq n_{closest-frame}$. For patient 49 cover 04, the closest-frame algorithm produces 155 clusters, a vast overestimate of the true number 33.

4 Clustering extensions to bridge discontinuities

The following section presents original proposals that extend the clustering above to deal with sperm track discontinuities.

4.1 Extending clustering methods through Hough transforms

When the sperms follow fairly linear paths, it follows to reason that if the tracks have similar rotations and velocities with respect to time, then they are likely to be the same sperm track. Equations for transforming the clusters to Hough-space can be found in the appendix E.1.

The Hough-transformed coordinates can be applied to any clustering algorithm to provide potentially accurate clusters by clustering in the Hough-space and converting back to Cartesian space. This method could be combined with DBSCAN with a low-tolerance distance parameter and high k value for optimal results. This method eliminates the issue of discontinuities where the track remains parallel across missing detections, but it may misclassify sperm tracks if there are intersections in the 2D Hough-space, which is caused by two parallel sperm tracks travelling at the same velocity. Derivation for the model complexity is included in the appendix D.2.

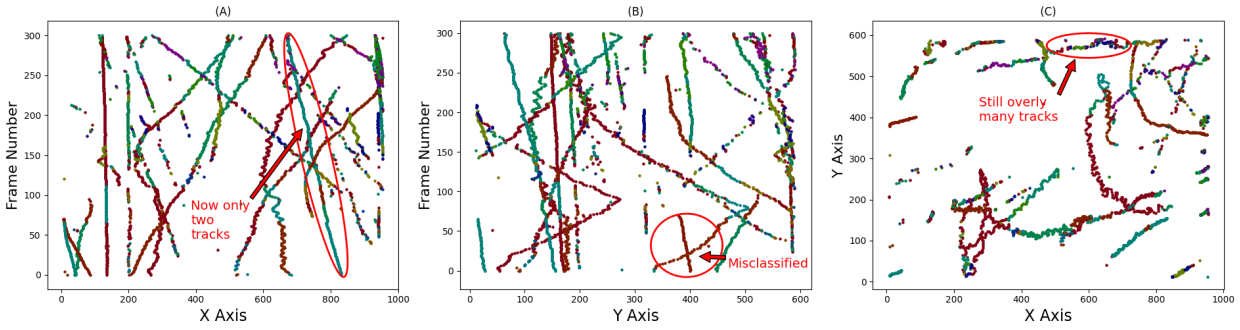


Figure 6: The extensions results through one iteration of closest-frame clustering, conversion to Hough-space, then finally clustered with DBSCAN are shown above. The results are not promising for the chaotic cover 04, since it now misclassifies points shown in Subfigure B which weren't wrongly tracked previously. Furthermore, it has failed to join any non-linear tracks as shown in Subfigure C. Thus, this extension is only effective in the scenario where the tracks are fairly linear, an example is portrayed in Subfigure A, where the algorithm has resolved 3 occluded clusters down to only 2.

4.2 Polynomial regression to infer missing detections

The principal issue with direct clustering is that the data will always contain major sperm track discontinuities and noise. Polynomial regression was implemented to predict the missing points between the discontinuities and BIC (Bayesian information criterion) was applied to determine the appropriate order of the polynomial regression. BIC was chosen over other methods like AIC (Akaike's Information Criteria) due to the uncertainty of the extrapolated data points, since BIC implements stronger penalties to additional parameters (higher orders) thereby decreasing the likelihood of overfitting [28]. The formulas for BIC and its implementation are included in the appendix E.2.

After clustering the data points for the first time using any method, the missing points between the discontinuities were predicted with BIC-enhanced polynomial regression. From the predicted points, an additional clustering algorithm can be implemented on the more continuous data. Finally, the surplus datapoints can be removed and the original state space has tracks which cannot be connected using any of the previous methods. The implementation of joining clusters is included in the appendix E.3.

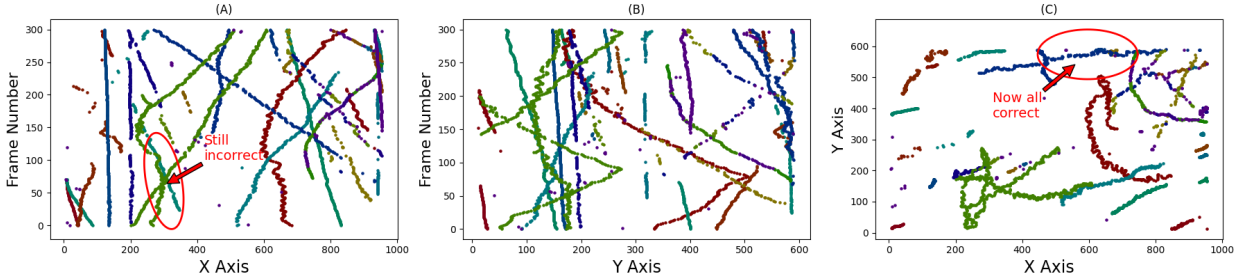


Figure 7: The most promising results for cover 04 were obtained with initially clustering with HDBSCAN, followed by BIC-polynomial regression extrapolation, then with a final second HDBSCAN clustering. It's clear in Subfigure A that the highlighted cluster is a combination of multiple tracks. This misclassification is caused HDBSCAN combining tracks where collision has occurred. In spite of this, many discontinuities are connected very well, such as shown in Subfigure C.

The success of predicted tracks produced by DBSCAN combined with BIC extrapolation, are shown on the left in Figure 8. By visually comparing the results to the raw video, most sperms were correctly clustered with minimal error. This method is limited when sperms cross paths, as seen in the bottom right corner of the hand-labelled Figure 8. There are two sperms that the model clustered as the same sperm, it was only after they crossed paths that it was correctly identified as a separate cluster. Given this limitation, all of the clustering models besides 'closest frame' perform poorly for more motile sperms, since this increases the probability of collisions and this is the only clustering algorithm to guarantee non overlapping sperm tracks. A possible reason this model had success for the first set of data could have been due to the sperm predominantly moving due to flow in the seminal fluid rather than the movement of the flagellum.

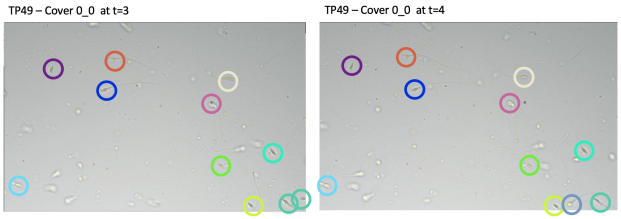


Figure 8: The results of the DBSCAN linear regression model for tp 49 cover 00 are shown above. The data is hand-labelled at time 3 seconds and 4 seconds to identify key limitation in the model.

5 Leading tracking algorithms

5.1 Joint Probabilistic Data Association Filter (JPDAF)

Unlike the previously considered methods, JPDAF is a multi-target tracking algorithm that explicitly attempts to associate tracks 'from a set of noisy and uncertain measurements' where 'observations generally include a set of spurious measurements' [29]. For this algorithm the main challenge is to 'estimate the state of an unknown and time-varying number of targets' [29]. A version of this algorithm was implemented using the MATLAB sensor fusion tracking toolbox; noise was added to reduce overfitting due to the sensitivity of this algorithm to changes in the state space. Given that the hyperparameters determine the model's uncertainty and state estimations, a subset of these were tuned by trial and error to fit the tracks of cover 04. Most notably a constant-turn-rate extended Kalman filter produced the best result as illustrated in Figure 9. Here JPDAF performs well overall; reconstructing short and discontinuous tracks to a higher fidelity, dealing with sharp cornering effectively and not terminating longer tracks prematurely, all of which has previously been problematic.

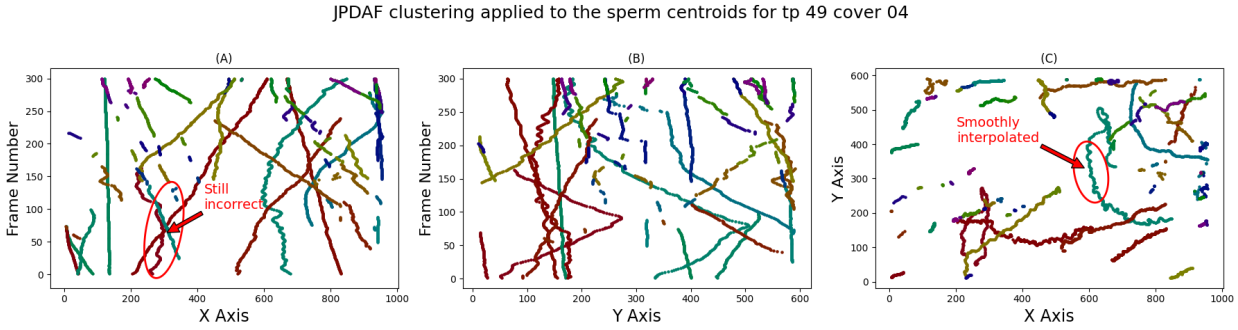


Figure 9: The results of JPDA with the following hyperparameters ("FilterInitializationFcn", @initctekf, "AssignmentThreshold",[200 Inf],"ConfirmationThreshold",[2 3],"DetectionProbability",0.999) are shown. The interpolation is incredibly precise and smooth regardless of the track, as shown by the chaotic track in Subfigure C, which is hugely beneficial. However, the algorithm often misclassifies intersections, similar to DBSCAN.

5.2 Particle tracking using IDL

The IDL particle tracking software, initially used to extract quantitative data from digitized video microscope images of colloidal suspensions, was developed by John C. Crocker and David G. Grier [30]. Here, a basic adaptation of this software in MATLAB, created by Daniel Blair and Eric Dufresne [31], was used to track sperm. The original software uses different image processing algorithms to determine the location of the particles based on various features. The key adaptions from the original algorithm are detailed in the appendix E.4.

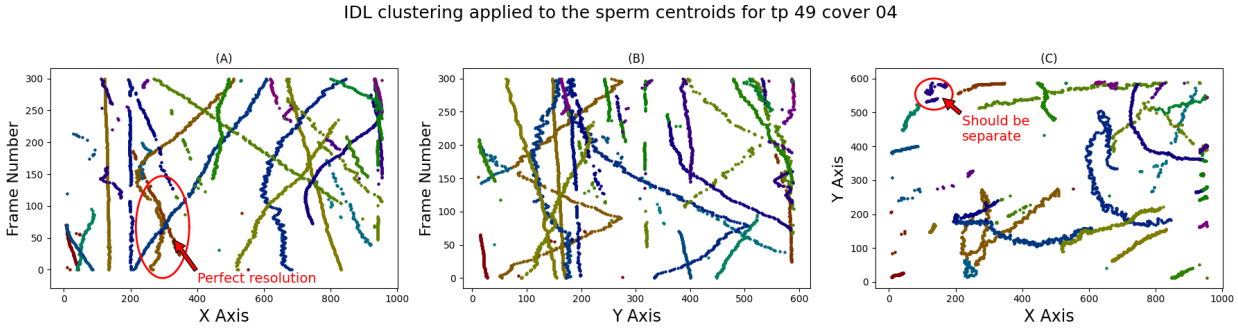


Figure 10: IDL tracking results using a maxdisp of 200 are shown. The algorithm is capable of correctly pairing discontinuities. Misclassifications are rare, and the algorithm easily resolves sperm collisions as intended, such as in Subfigure A. IDL tracking does not support grouping noise, so unfortunately, additional approaches, like DBSCAN or noise filters must be implemented to distinguish whether or not a track is noise or sperm.

5.3 Track-Oriented Multi-Hypothesis Tracker (MHT)

The MHT approach to tracking sperm allows further and more precise inspection into potential tracks [32]. The tracker initializes tracks based on the points for the initial time. Then, the tracker will attempt to assign the set of points of the next time step to existing tracks. The points having not been assigned will initialize new tracks. However, unlike the other methods mentioned, multiple hypotheses are allowed regarding the assignments of detections to tracks. These assignments create branches within the assigned tracks. The branches are then given an initial scoring; the lowest scoring tracks are cut off. Then, clusters of the incompatible branches are generated (branches containing the same detection). Following this, the global hypotheses of compatible branches are formulated and scored. All branches are then scored based on their existence in the global hypotheses, with the low scoring branches being cut off.

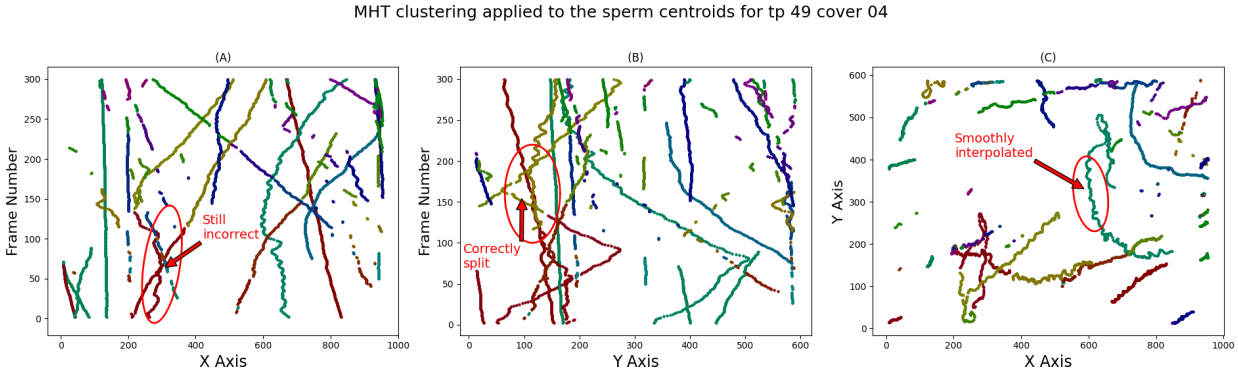


Figure 11: The results of MHT with hyperparameters ("FilterInitializationFcn", @initctekf, "DetectionProbability",0.999) are shown. Similar to JPDAF, MHT smoothly interpolates between discontinuities. Additionally, MHT produces fewer misclassifications, although the number of predicted tracks has increased by 30 to account for this.

6 Discussion

Devising a suitable tracking algorithm first requires a thorough examination of the considered sperm detection algorithms. For this, different performance metrics were implemented and the computational complexity was considered alongside, since efficiency is key in improving on hand-labelling. Furthermore, it was important to consider the accuracy of the detection algorithms from the original five second sperm video samples.

Detection rates from the videos were consistently high and almost all sperm were detected in at least one frame across the video. The exceptions to this were primarily sperm that were out of focus and thus blurrier in the footage. The cases where re-detection occurred frequently mainly consisted of motile and immobile sperm at the edge of the frame, motionless sperm that are abnormally shaped and agile sperm which move rapidly and take sharp deviations often dipping in and out of the camera's focus. The difficulty of automatically diagnosing these detection failures and correctly linking sperm tracks is compounded by the issue of the less frequent, inaccurate detections of the system due to noise.

Substances like cellular matter and debris are incorrectly identified as sperm, illustrated in Figure 12. These false positives can last anything from a single frame to being a systematic error: e.g. the image on the right in Figure 12. If a noisy signal occurs only in a few frames, it can be difficult to distinguish from an agile sperm whose track has recently been lost. A common cause for false positives is when a sperm's tail brushes up against or aligns with debris of a certain size and shape as shown in Figure 13. When this occurs, the illusion of two separate sperm is given: tricking the detection system into misrepresenting the debris. Therefore, a proposal of this report is that sperm tail dynamics or information is incorporated into the sperm classification algorithm to reduce noise.

Figure 12: The detected noise in video samples includes other cellular matter (left: patient 57, cover 09, frame 93) and debris (middle: patient 57, cover 09, frame 177) , (right: patient 49, cover 12, frame 1).

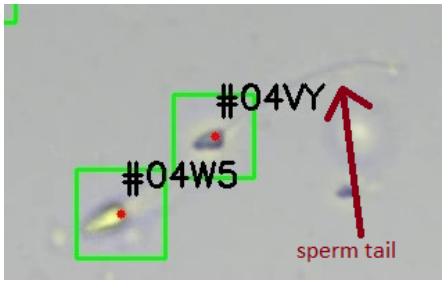


Figure 13: Presented is an exemplar image of a false positive detection. This detection is resulting from a sperm tail alignment with debris from patient 57, cover 06, frame 215. The sperm head and debris is bounded by the points labelled #04W5 and #04VY respectively.

Another issue identified from the footage is that localised flow caused by fluid sheering forces the immotile sperm to drift artificially. An example of this is cover 00 of patient 49, as shown in Figure 14, where inactive sperm move without beating their tails. One such case is circled in yellow. Consequently, a tracking system may incorrectly skew its evaluation of sperm motility based on their motion with respect to localised flows. Furthermore, such flows appear to increase the rate of false positive detections as exemplified by cover 0 of patient 49 suggesting that the detection algorithm is biased towards moving objects. A recommendation is that samples are prepared in a way that ensures no causes of flow are present e.g. air bubbles or temperature gradients.

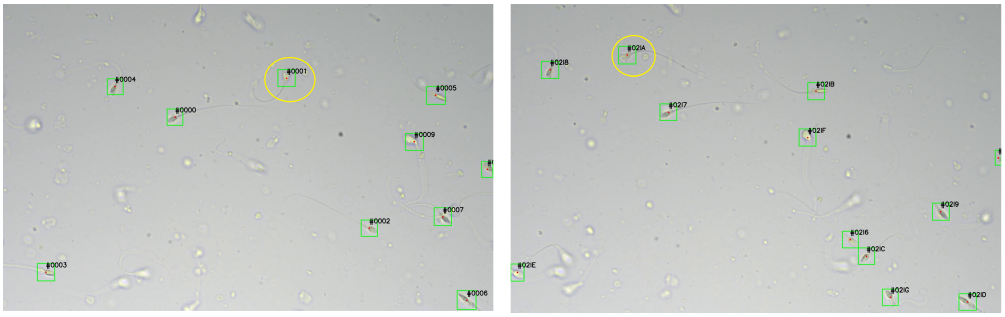


Figure 14: Images of patient 49, cover 00, (left: frame 1, right: frame 301) highlighting one of the sperms (circled in yellow) whose movement is caused exclusively by a localised fluid flow (it does not move its flagellum).

It must be noted that whilst the intention was to categorise any points added by the trackers as "interpolated", both trackers that implement this (JPDAF and MHT) would also smooth tracks to a degree much greater than typical distances between sperms. Therefore, no basic heuristic could be implemented that is provably robust to false positive classification.

The hyperparameter space for JDPAF and MHT were not fully or systematically explored due to their size and complexity. Since it is likely that there exist configurations better adapted to complex sperm interactions and the hyperameters have not been tuned across all covers, it is recommended that further research explores this configuration space with an appropriate loss function. Another consideration that could result in significant improvement in the tracking algorithms is the use of data prepossessing to denoise the data. Additional steps can be taken to clean the detections of false positives thus restricting the notable influence of such points on the tracks reconstructed by the algorithms.

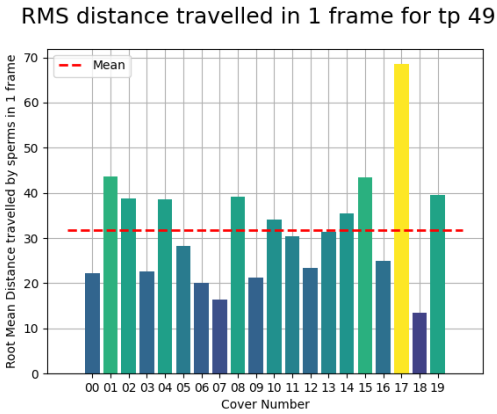
A comparison Table 1 is shown below to easily compare the accuracy against the computational complexity of each existing model. The standard MOTA metric [21] was only implemented for tp 49, cover 04, therefore the results may vary wildly across different samples. The MOTA metric is an industry standard percentage accuracy, with its derivation in the appendix D.5. It is impossible to extend this across more samples, since this requires hand-clustering to obtain the ground-truth. A proposal metric 'U' measures cluster distribution and its derivation is included in the appendix D.4. A lower 'U' value corresponds to a more accurate tracking distribution, and a lower 'U' standard deviation corresponds to a more robust/generalising method. A similar proposal metric named 'convex hull density' was implemented. This was created by obtaining the density of the convex hull for each cluster, which should represent how few discontinuities an algorithm contains, because sparse clusters will tend to have a high density. Lastly, the time and space complexities alongside the mean runtimes were included to validate the cost of accuracy against complexity. It is important to understand that it is mathematically impossible to construct a perfect performance metric without having a perfect tracking algorithm, since one could optimise for the performance metric and solve the problem.

Algorithm	**U Mean**	**U Standard Deviation**	**Convex Hull Density Mean**	**MOTA**	**Time Complexity**	**Space Complexity**	**Mean runtime per 300 frames (s)**
None (1 cluster)	16.335	8.62	9,984	-	$O(1)$	$O(1)$	0.029
Ground Truth (TP 49, COVER 04)	0.612	0.00	20,641	100%	-	-	~18,000
K-means	0.941	0.65	3089	73.72%	$O(nk)$	$O(n)$	0.239
DBSCAN	0.805	0.23	697	74.30%	$O(n^2) \mid O(nlog(n))$	$O(n) \mid O(n^2)$	0.06
HDBSCAN	0.818	0.16	720	79.87%	$O(n^2)$	O(n)	0.288
GMM	0.608	0.18	484	73.37%	$O(n)$	$O(n^2)$	0.443
Closest Frame	0.865	0.14	38,726	80.97%	$O(n)$	O(n)	0.399
Optimal Hough Transform	1.250	0.32	30,980	84.57%	$O(nn)$	$O(n^2)$	0.416
Optimal Linear Extrapolation	0.971	0.57	3608	88.54%	$O(n^2)$	$O(n^2)$	0.386
Optimal BIC-polynomial Extrapolation	0.843	0.17	443	89.76%	$O(n^2)$	$O(n^2)$	0.452
IDL Particle Tracking	0.567	0.13	155,163	98.70%	$O(n^2)$	$O(n^2)$	3.744
JPDAF	0.74	0.10	578	78.57%	$O(n^3) \leq O(JPDAF) \leq O(n^2 4^{2n})$	$O(n^2)$	0.071
MHT	0.836	0.08	255	86.78%	$O(JPDA) \leq O(MHT) \leq O((n^2)!))$	$O(n^2)$	0.188

Table 1: The cross-metric results are for every considered algorithm. The ground-truth was only obtained for patient 49, cover 04, therefore the MOTA metric was only calculated for this one sample, so results may vary across samples. It is worth noting that the GMM algorithm 'cheats' the 'U' mean, standard deviation and the convex hull density. This is because GMM effectively minimises the distributions of the samples. Missing from the table is the popular Open CASA [33] algorithm: it was not considered due to the integration difficulty caused by unfriendly interfaces.

Finally, measuring the approximate mean motility of a sperm sample was a corollary of the 'closest-frame' algorithm. The root mean squared (RMS) speeds for the sperm should be calculated from the single-frame distances matrix. Consistent with maxwell-boltzmann theory for particle energies, the greater the RMS is, the greater the 'motility', which is a measure of the mean sperm velocities in the sample, as portrayed below in Figure 15.

(a)



(b)

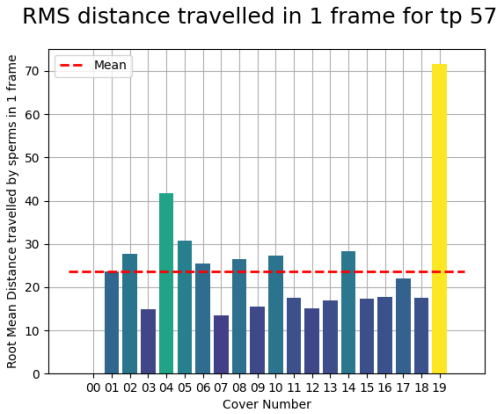


Figure 15: The root mean squared distance travelled by sperms in 1 frame for the recorded datapoints is shown in the graph for each different cover in each test patient, 49 and 57. The results for tps 49 and 57 are different subjects' sperm samples, with each cover being a subsample of each sperm sample. Its clear that tp 49 has more motile sperm, with a greater mean RMS speed of 31.79 per frame compared to that of tp 57: 23.54.

7 Conclusion

The final process for determining how to apply each different sub-algorithm can be seen in the appendix C. Basic clustering algorithms: K-means and GMM always perform poorly for all samples, since they assume spheroid clusters, whereas sperm tracks are continuous paths. DBSCAN better separates clusters, but misclassifies sperms when there are sperm collisions and cannot bridge discontinuities. HDBSCAN cannot bridge discontinuities either, but it produces fewer misclassifications. Closest-frame clustering never misclassifies sperms, but it produces the greatest possible number of discontinuities. Two extensions to combine clusters with discontinuities are suggested: BIC extrapolation clustering and Hough transform clustering. Clustering the Hough-space and converting back does not ensure no misclassifications and it is inaccurate for non-linear data, but it is able to effectively reduce the number of clusters in the highly linear case. BIC-based extrapolation improves on this and is able to effectively bridge clusters, however any unintended collisions between extrapolated data may combine two different sperms into one cluster.

Traditional leading methods for tracking are MHT, JPDA, IDL and of course, hand-labelling. Hand-labelling should be rejected as it is preferable to apply more inaccurate methods, since it requires around 5 hours of dedicated analysis by a professional, per singular cover of only 300 frames. Both JPDAF and MHT produce commendable results by tracking the general behaviour of sperms and interpolating points for track smoothing. However, JPDAF and MHT are infeasible; scale poorly. IDL provides the best performance and is highly suited because of: effective treatment of dense track coalescing, strong ability to resolves discontinuous tracks amidst noise, minimal hyperparameter tuning, it has the lowest U mean score and comparable time and space complexity to clustering approaches.

Comparing all of the proposed algorithms, the optimal performance metrics pertain to IDL tracking, followed swiftly by BIC-polynomial extrapolation, based upon HDBSCAN clustering. In spite of the excellent results that IDL tracking produces, BIC-polynomial tracking runs autonomously without hyperparameter tuning, thus may be potentially better suited for technicians at the marginal cost of accuracy.

References

[1] Sung-Yang Wei, Hsuan-Hao Chao, Han-Ping Huang, Chang Francis Hsu, Sheng-Hsiang Li, and Long Hsu. A collective tracking method for preliminary sperm analysis. *BioMedical Engineering OnLine*, 18(1), nov 2019.

[2] Cendrine Geoffroy-Siraudin, Anderson Dieudonné Loundou, Fanny Romain, Vincent Achard, Blandine Courbière, Marie-Hélène Perrard, Philippe Durand, and Marie-Roberte Guichaoua. Decline of semen quality among 10 932 males consulting for couple infertility over a 20-year period in marseille, france. *Asian Journal of Andrology*, 14(4):584–590, apr 2012.

[3] SK Adiga, V Jayaraman, G Kalthur, D Upadhya, and P Kumar. Declining semen quality among south indian infertile men: A retrospective study. *Journal of Human Reproductive Sciences*, 1(1):15, 2008.

[4] Hagai Levine, Niels Jørgensen, Anderson Martino-Andrade, Jaime Mendiola, Dan Weksler-Derri, Irina Mindlis, Rachel Pinotti, and Shanna H Swan. Temporal trends in sperm count: a systematic review and meta-regression analysis. *Human Reproduction Update*, 23(6):646–659, jul 2017.

[5] Niels Jørgensen, Ulla Nordström Joensen, Tina Kold Jensen, Martin Blomberg Jensen, Kristian Almstrup, Inge Ahlmann Olesen, Anders Juul, Anna-Maria Andersson, Elisabeth Carlsen, Jørgen Holm Petersen, Jorma Toppari, and Niels E Skakkebæk. Human semen quality in the new millennium: a prospective cross-sectional population-based study of 4867 men. *BMJ Open*, 2(4):e000990, 2012.

[6] Yash S. Khandwala, Chiyuan A. Zhang, Ying Lu, and Michael L. Eisenberg. The age of fathers in the USA is rising: an analysis of 168 867 480 births from 1972 to 2015. *Human Reproduction*, 32(10):2110–2116, aug 2017.

[7] Statistics Norway. Average age of parents at child's birth (sy 72) [online]. Available from: `https://www.ssb.no/260132/average-age-of-parents-at-childs-birth-sy-72`, N/A. [Last Accessed: 20/02/2021].

[8] Office for National Statistics. Births by parents' characteristics [online]. Available from: `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datasets/birthsbyparentscharacteristics`, nov 2020. [Last Accessed: 20/02/2021].

[9] E. Levitas, E. Lunenfeld, N. Weisz, M. Friger, and G. Potashnik. Relationship between age and semen parameters in men with normal sperm concentration: analysis of 6022 semen samples. *Andrologia*, 39(2):45–50, apr 2007.

[10] Sheri L. Johnson, Jessica Dunleavy, Neil J. Gemmell, and Shinichi Nakagawa. Consistent age-dependent declines in human semen quality: A systematic review and meta-analysis. *Ageing Research Reviews*, 19:22–33, jan 2015.

[11] Rosa Isabel Molina, Ana Carolina Martini, Andrea Tissera, José Olmedo, Daniel Senestrari, Marta Fiol de Cuneo, and Rubén Daniel Ruiz. Semen quality and aging: analysis of 9.168 samples in cordoba. argentina. *Archivos espanoles de urologia*, 63:214–222, April 2010.

[12] World Health Organisation. Who laboratory manual for the examination and processing of human semen [online]. Available from: `https://www.who.int/publications/i/item/9789241547789`, 2010. [Last Accessed: 1/03/2021].

[13] Leonardo F. Urbano, Puneet Masson, Matthew VerMilyea, and Moshe Kam. Automatic tracking and motility analysis of human sperm in time-lapse images. *IEEE Transactions on Medical Imaging*, 36(3):792–801, mar 2017.

[14] Steven A. Hicks, Jorunn M. Andersen, Oliwia Witczak, Vajira Thambawita, Pl Halvorsen, Hugo L. Hammer, Trine B. Haugen, and Michael A. Riegler. Machine learning-based analysis of sperm videos and participant data for male fertility prediction. *Scientific Reports*, 9(1), nov 2019.

[15] Felix Ruhnow, David Zwicker, and Stefan Diez. Tracking single particles and elongated filaments with nanometer precision. *Biophysical Journal*, 100(11):2820–2828, jun 2011.

[16] Michael J Saxton. Single-particle tracking: connecting the dots. *Nature Methods*, 5(8):671–672, aug 2008.

[17] Yannis Kalaidzidis. Multiple objects tracking in fluorescence microscopy. *Journal of Mathematical Biology*, 58(1-2):57–80, may 2008.

[18] I. Smal, K. Draegestein, N. Galjart, W. Niessen, and E. Meijering. Particle filtering for multiple object tracking in dynamic fluorescence microscopy images: Application to microtubule growth analysis. *IEEE Transactions on Medical Imaging*, 27(6):789–804, jun 2008.

[19] I.F. Sbalzarini and P. Koumoutsakos. Feature point tracking and trajectory analysis for video imaging in cell biology. *Journal of Structural Biology*, 151(2):182–195, aug 2005.

[20] B. Ristic, B. Vo, D. Clark, and B. Vo. A metric for performance evaluation of multi-target tracking algorithms. *IEEE Transactions on Signal Processing*, 59(7):3452–3457, 2011.

[21] Keni Bernardin, Alexander Elbs, and Rainer Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. *Proceedings of IEEE International Workshop on Visual Surveillance*, 01 2006.

[22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[23] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer Berlin Heidelberg, 2013.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[25] Imad Dabbura. K-means clustering: Algorithm, applications, evaluation methods, and drawbacks [online]. Available from: `https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a`, 2018. [Last Accessed: 19/02/2021].

[26] Steve Astels Leland McInnes, John Healy. Combining hdbscan with dbscan [online]. Available from: `https://hdbscan.readthedocs.io/en/latest/how_to_use_epsilon.html`, 2016. [Last Accessed: 19/02/2021].

[27] Eyal Trebelsi. Comprehensive guide to approximate nearest neighbors algorithms [online]. Available from: `https://towardsdatascience.com/comprehensive-guide-to-approximate-nearest-neighbors-algorithms-8b94f057d6b6`, 2020. [Last Accessed: 19/02/2021].

[28] John J. Dziak; Donna L. Coffman; Stephanie T. Lanza; Runze Li. Aic vs bic differences between aic and bic methods [online]. Available from: `https://www.methodology.psu.edu/resources/AIC-vs-BIC/`, 2019. [Last Accessed: 24/02/2021].

[29] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

[30] John C. Crocker and David G. Grier. Methods of digital video microscopy for colloidal studies. *Journal of Colloid and Interface Science*, 179(1):298–310, 1996.

[31] Daniel Blair and Eric Dufresne. Matlab particle tracking code repository [online]. Particle-tracking code available at http://physics.georgetown.edu/matlab/. [Last Accessed: 28/02/2021].

[32] John R. Werthmann. Step-by-step description of a computationally efficient version of multiple hypothesis tracking. In Oliver E. Drummond, editor, *Signal and Data Processing of Small Targets 1992*, volume 1698 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 288–300, August 1992.

[33] Opencasa: A new open-source and scalable tool for sperm quality analysis [online]. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006691#sec002. [Last Accessed: 24/12/2020].

[34] David Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector [online]. Available from: https://doi.org/10.5281/zenodo.3240529, June 2019.

[35] Martin Simonovsky. Ellipse detection using 1-d hough transform [online]. Available from: https://fr.mathworks.com/matlabcentral/fileexchange/33970-ellipse-detection-using-1d-hough-transform?s_tid=srchtitle. [Last Accessed: 10/02/2021].

[36] R. Fisher; S. Perkins; A. Walker; E. Wolfart. Hough transform [online]. Available from: https://homepages.inf.ed.ac.uk/rbf/HIPR2/hough.htm. [Last Accessed: 02/02/2021].

A EDI Statement

EDI issues are systemic and are becoming increasingly frequent in modern systems that integrate automated machine learning. Any AI system that interacts with people, or uses data in any way connected with people needs to be checked carefully in order to avoid any cases of discrimination or implicit bias. Measures should be taken to ensure that quality control and due diligence are followed throughout the data pipeline. Procedures such as ensuring data sources are accurate and reliable, follow the same data collection methodology, representative of the population and of sufficient quantity can markedly reduce cases of discrimination. Similarly, it is vital that the principals behind the systems design - its algorithms and metrics included- are absent of favouritism to subsets of the data (weighting all samples equally) or certain features. Such a system should be supported by a phase of extensive fairness testing of the code base to ensure no causes of human error or prejudice are present including in the data prepossessing stage. Furthermore, the algorithm should be exposed in both training and testing to a diverse dataset. After development, the appropriate system monitoring should be undertaken to provide outcome fairness and implementation fairness to avoid 'discriminatory or inequitable impacts on the lives of the people they affect' and 'deployed by users sufficiently trained to implement them responsibly and without bias' [34].

More specifically to sperm analysis, consent is necessary at all stages of sample analysis. If a subject no longer provides consent then their sample should be destroyed in the appropriate manner and their data erased. An important consideration is informing the subject beforehand to exactly how their data will be used and processed, which has been adhered to. If the data processing requirements then change, and the subject does not agree/is not informed this is regarded as a privacy breach and illegal under data privacy laws. The information of the subjects of this report is anonymous and treated equally, but if later published with other information a high level of k-anonymity should be maintained to ensure that individuals aren't re-identified as this could cause distress and carry other implications.

Finally, this report makes no ethical case; it does not support/condone or oppose the use of genetic screening or other such process that could be used to discriminate, in artificial insemination or in-vitro fertilisation, against spermatozoa that could result in a genetic disability or any other basis of prejudice.

B Covid mitigation statement

The global pandemic is an evident and primary obstacle for current group projects. The pandemic has had obvious implications in challenges related to communication, cooperation and personal difficulties. All things considered however, this group has effectively used online channels to overcome such difficulties and has not had to deal with issues such as timetable differences. Overall, the team has adapted well and collaborated in a successfully manner under the circumstances. No other noteworthy impacts have been identified in relation to the pandemic.

C Data analysis and workflow

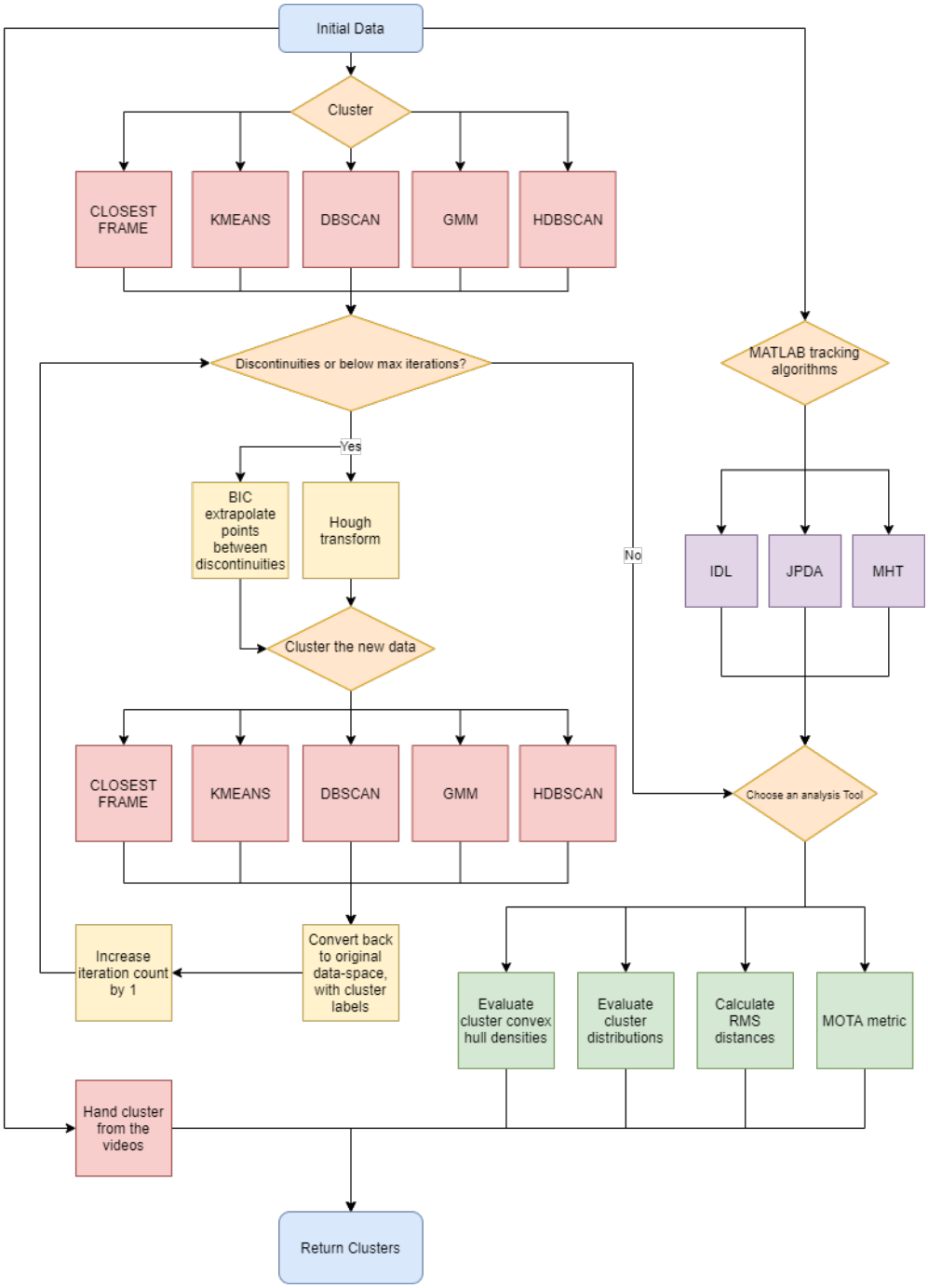


Figure 16: The flowchart for clustering the data above indicates the simple clustering methods, the iterative extensions and the different analysis tools that can be interchanged to evaluate the tracking algorithms. Also included with evaluations are the different motility and classification techniques that can be applied.

D Proofs and derivations

D.1 Closest frame clustering complexity

The time and space complexities for this algorithm are approximately $O(n_f n_s^2)$ where n_f is the number of frames and n_s is the average number of sperms in each frame. As the number of frames and the number of datapoints n increases, the time and space complexities both tend to complexity $O(n_f n_s^2) \rightarrow O(n n_s)$. For the training data, n_s is always less than 50, as expected for sperm samples. Therefore, n_s can be treated as fairly constant, thus the time and space complexities tend to $O(n_f n_s^2) \rightarrow O(n)$ as n becomes significantly large.

D.2 Hough transform complexity

A naive approach would be to apply the Hough transform to all pairs of points, which would have time complexity $O(n^2)$ and space complexity $O(n^2)$, where n is the number of datapoints. Quadratic time complexity is infeasible and does not scale well. Instead, by treating each cluster input as a sequence of points, ordered by frame number, it allows the time and space complexities for converting from Cartesian to Hough-space to be approximately $O(m^2 k)$ where m is the mean number of points in each cluster and k is the total number of clusters. As the total number of datapoints n increases, mk tends to n, thus the complexity $O(m^2 k) \rightarrow O(mn)$. The mean number of points in each cluster, m is subject to increase slowly as n increases, therefore the final time and space complexities are between linear and quadratic: $O(n) << O(mn) << O(n^2)$.

D.3 Deriving the sperm head orientation

So far, the tracking has been done relying solely on the detection based data annotations. However, there is additional data to exploit through the sperm videos. The orientation of a sperm head can lead to key information in predicting a sperm's next position. Knowing the orientations limits the possible range of positions in the next frame by excluding points that are highly improbable. To obtain this information, a MATLAB program for ellipse detection in images was used [35]. This information is particularly valuable at track intersections or collisions to distinguish sperms.

The program first converts the bounding box image to a canny image, where only the pixels over a certain intensity threshold are kept and it produces an edge detection image. Then, using the canny image of the bounding box, the best fitting ellipse is determined by examining all possible major axes (all pairs of points) and getting the minor axis using a Hough transformation [36]. The effective run-time of this program can be reduced by specifying parameters and constraints to the problem. By constraining the major axis of the ellipse by using sperm head length constraints the number of potential ellipses decreases drastically.

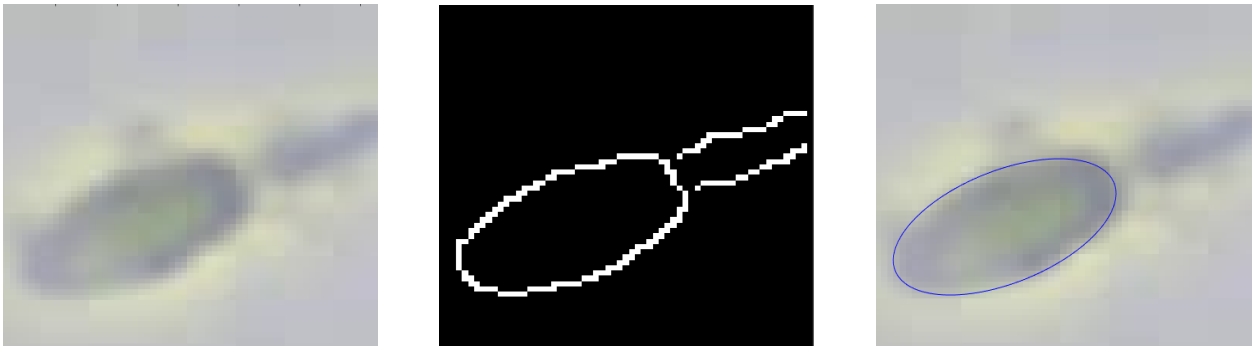


Figure 17: Ellipse detection for sperm head stages : original bounding box (left), canny edge detection image (middle) and fitted ellipse (right). The program returns the center coordinates, the major and minor axis lengths, the angle between the major axis and the x-axis as well as a score based on the accuracy of the fitted ellipse with respect to the edge points detected.

By iterating the algorithm over all the bounding boxes in each frame of a video, the angle between the x-axis and the sperm head can be recorded for every sperm instance. However, this angle is not necessarily representative of the direction the sperm head is pointing; it just indicates the alignment of the sperm head with this angle.

To find the directed orientation, further analysis on the bounding box image is necessary. By locating the position of the sperm's tail in relation to the ellipse and coupling this with the orientation, the sperm head's direction can be approximated. Two regions of interest can be defined. Using the ellipse's extremity points, the potential areas containing the tail can be narrowed down to two rectangular regions represented in Figure 18. The tail is assumed to be in the region of interest having the highest pixel intensity using the canny image of the bounding box. The canny edge detection is set to an optimal threshold of 0.5 so that noise is minimised and the edge pixels displayed are only representative of the head and tail of a sperm. The tail's location coupled with the raw head orientation is enough to define the direction in which the sperm is going by extending the major axis into the highest pixel intensity region of interest and selecting a point to be the tail's position. The vector from the tail's position to the center of the ellipse represents the sperm head's direction. Unfortunately, using this information led to worse results in Section F.

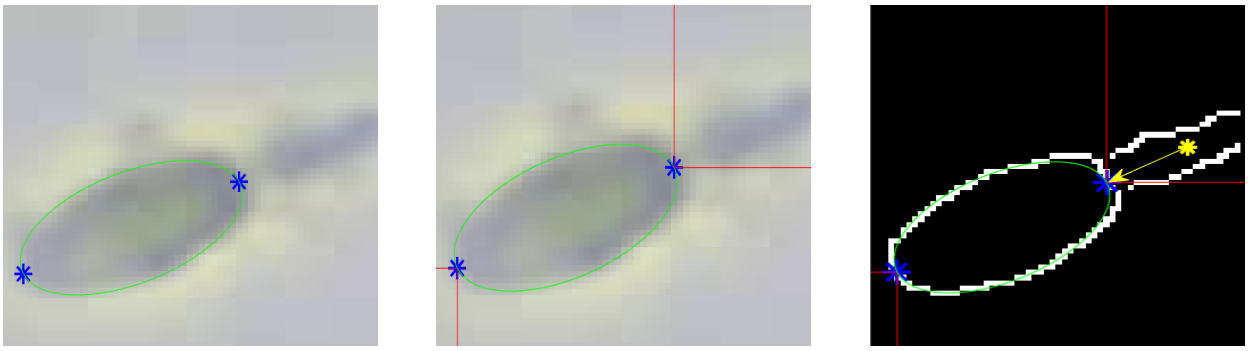


Figure 18: Locating the sperm's tail: finding the ellipse extremities (left), defining the regions of interest regarding the sperm tail's position (middle), selecting the region of interest based on pixel intensity to approximate the tail's location (right).

D.4 Cluster distribution: 'U' value derivation

Fortunately, the video contains a set number of frames, n_f, and the majority of sperms should remain within the camera's field of view. Therefore, it is possible to generate a performance metric for each cluster's number of points n_c compared to n_f for all k clusters. The 'U' value can be defined as the following:

$$U = \frac{1}{k\, n_f} \sum_{cluster \in clusters} |n_f - n_c| \quad . \tag{1}$$

D.5 Multiple object tracking accuracy

Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) are performance metrics for multiple object trackers[21]. First, object-hypothesis correspondences are made for every frame based on their distance.

MOTA measures a tracker's performance at keeping accurate trajectories, independent of its precision in estimating object positions.

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{2}$$

where m_t, fp_t and mme_t are the number of misses, of false positives and of mismatches respectively for time t and g is the number of matches at time t.

The MOTA metric can be used as an indication of how well a tracker performs. However, the implementation seems to give abnormally high accuracy scores to the trackers overall and should be modified for better scaling.

E Additional formulae

E.1 Hough transform formulae and their implementation

The 2D linear Hough transform algorithm [36] estimates the two parameters $(\mathbf{r}, \theta)$ that define straight lines from 2D Cartesian space to 2D Hough space. For pairs of datapoints $\mathbf{X_j} = (x_j, y_j)$ and $\mathbf{X_i} = (x_i, y_i)$ then equations 3, 4 and 5 can be defined iteratively for the given pairs i and j for $i \neq j$.

$$\mathbf{v_i} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} (\mathbf{X_j} - \mathbf{X_i})^T \tag{3}$$

$$\text{for } \mathbf{v_i} \neq \mathbf{0},\ r_i = \mathbf{x_i} \cdot \frac{\mathbf{v_i}}{|\mathbf{v_i}|} \tag{4}$$

$$\theta_i = |\arg \mathbf{v_i}| \tag{5}$$

The angles are unique to their respective dimensions, $\mathbf{x}$ and $\mathbf{y}$, the mean, presented by equation 6, will transform them from planar space to 3D space. The number of dimensions is reduced from four to three, which will increase the efficiency of any clustering algorithms applied to the Hough-transformed data.

$$\bar{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta_x} + \boldsymbol{\theta_y}}{2} \tag{6}$$

E.2 BIC formula and its implementation

BIC was used infer extrapolated data points between the discontinuities. BIC is formally defined as [28]:

$$BIC = n \ln \left(\frac{RSS}{n} \right) + K \ln (n) \tag{7}$$

where:

- RSS is the residual sum of squares
- n is the sample size
- K the number of parameters used estimated by the model (the order of the polynomial)

BIC is used to find the optimal order for polynomial regression by evaluating the BIC value for each K, where $1 \leq K \leq 12$. The K value chosen is the K before an increase of BIC (ie, where $BIC_k - BIC_{k-1} > 0$). An example is illustrated in Figure 19.

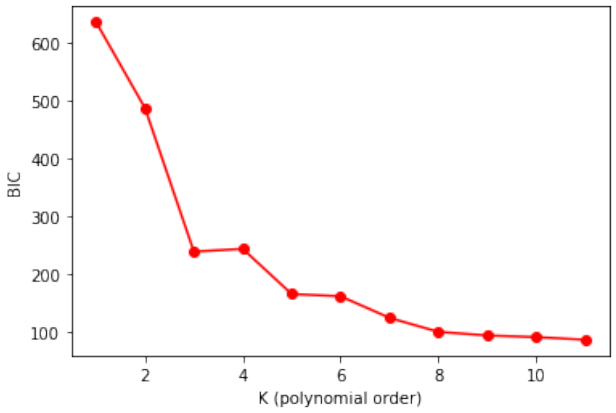


Figure 19: The graph illustrates a rapid decrease of the BIC value for $K = 1$ to $K = 3$ going from 650 to 250 respectively for a fixed sample size. For $K = 4$, the BIC value goes up to 265 approximately. Finally for $K > 4$ the graph continues to decrease but in a slower manner. Since the BIC value increases from $K = 3$ to $K = 4$, the algorithm determines that a 3rd order polynomial is best to fit and predict the missing data.

Two independent BIC assisted polynomial regressions are applied to predict the missing data points; both the x and y dimensions. An example is illustrated below in Figure 20, where the x and y coordinate have been predicted. Note that the polynomial regression predicted for the x-coordinate is of second order, whereas the y-coordinate is only of first order, so the order of each coordinate is independent.

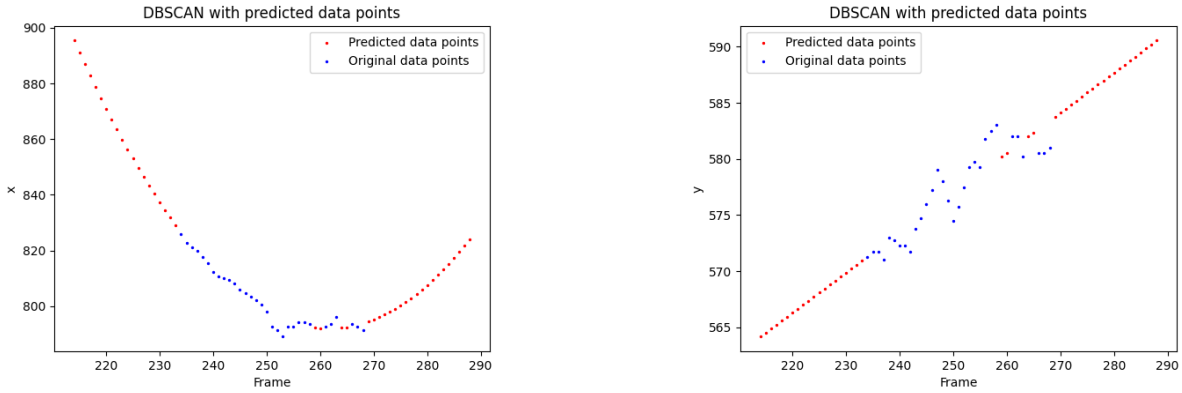


Figure 20: Illustrated are the predicted data points for a linear and a nonlinear regression, from a DBSCAN clustering of tp 49, cover 04.

E.3 An example of reclustering with the extrapolated data points

Joining discontinued clusters together can be done by using the extrapolated data points found in the previously in the appendix E.2. An example of joining discontinued clusters with the infer data points is illustrated in Figure 21.

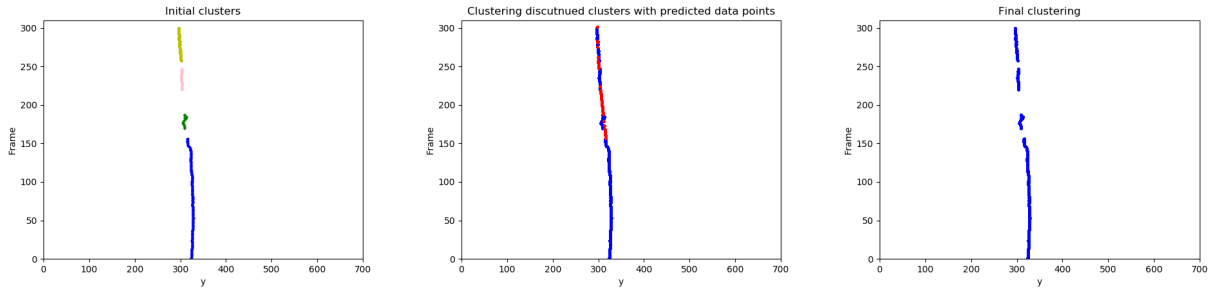


Figure 21: Illustrated is the process of joining discontinued cluster, where four disconnected clusters were joined to one cluster. The initial four discontinued clusters is illustrated in the left image with four different colours (blue, green, pink and yellow), the image in the middle illustrates the four clusters joined together in blue with the the extrapolated data points between the clusters in red, finally, the image on the right illustrates only one cluster with the surplus datapoints removed.

E.4 IDL particle tracking key adaptions

The adaptation used does not deal with the image processing aspect of the software; instead, the locations and corresponding detection times were passed through as a list. These locations are then linked into trajectories.

The algorithm's adaptation links the positions of the n sperms at time t and the m sperms at time $t + step$ (here, all steps are one frame) by considering all possible identifications of the n old positions with the new m positions. It chooses the identification set which minimises the total squared displacement. The identifications which don't associate a new position within the user-specified maximum displacement per time step $maxdisp$ will penalize the total squared displacement by $maxdisp^2$.

F Further results using orientation data

The results in Figure 22 show that adding the orientation data as the third dimension leads to a decrease in accuracy as it creates many more tracks. The method to incorporate the data of the sperm head's direction is not optimal; ideally a filter would be created to use the orientation data and assign its impact on the cost matrix. Further research is recommended to evaluate the utility of sperm head orientation for tracking.

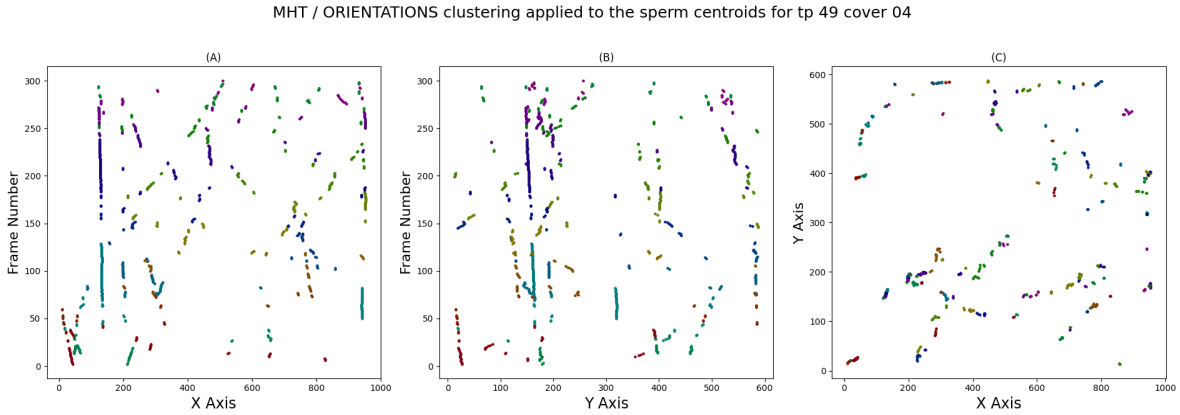


Figure 22: Results for MHT tracker are displayed, with default hyperparameters using sperm orientation as the third dimension for tracking. The results are very poor and it is clear that there are many discontinuities caused by inconsistent data removal by the algorithm.