Tracking of Spermatozoa by YOLOv5 Detection and StrongSORT with OSNet Tracker

Martin Kosela[1,*], Jakub Aszyk[2], Mateusz Jarek, Jakub Klimek[3] and Tomasz Prokop[3]

[1] *intelekt.us*
[2] *Poznan University of Technology, Poland*
[3] *University of Warsaw, Poland*

Abstract

This paper describes the solution to *Medical Multimedia Task: Transparent Tracking of Spermatozoa* using YOLOv5 object detection and StrongSORT with OSNet tracking algorithms. Using these techniques and carefully adjusted parameters and custom methods for motility analysis we achieve the tracking accuracy of $HOTA_{AUC} = 0.283$ for normal sperm detection task. Furthermore, we propose a novel method for determining the sperm motility by comparing average cell velocity at different sampling rates.

1. Introduction

The analysis of spermatozoa motility in human semen samples is an important task in male fertility treatment. Few attempts were made to solve this task automatically using computer vision algorithms, however, none of them showed high enough accuracy and reliability. Hence, this analysis is still being performed manually by qualified technicians which implies high test costs and inaccuracy due to interpersonal variety in technicians performance.

In this work, we solve the problem of sperm motility measurement in 3 stages:

- Stage 1: Detecting the sperm cells on each frame separately.
- Stage 2: Tracking the sperm cells by assigning them a unique id that does not change throughout the video.
- Stage 3: Computing and analyzing the sperm cells velocity.

Stages 1 and 2 correspond to sub-task 1 from the *Medical Multimedia Task: Transparent Tracking of Spermatozoa* while stage 3 corresponds to sub-tasks 2 and 3. The full details of the task and sub-tasks are described in [1]. In the following sections we briefly describe the approach, challenges and results for each stage.

2. Related Work

Object detection is a well-studied computer vision task with numerous good solutions using deep neural networks. We decided to use the state of the art approach YOLO, first described by Redmon et al. in 2015[2], which offers high accuracy while preserving fast inference time. Among many existing solutions to the object tracking problems [3, 4, 5, 6, 7], we decided to use StrongSORT, first described by Du et al. in 2022 [4] with OSNet by Zhou et al. from 2019 [5].

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

[*] Corresponding author.

martin@intelekt.us (M. Kosela)

CEUR Workshop Proceedings (CEUR-WS.org)

Previous studies of sperm analysis [8, 9, 10] show great potential of machine learning approach to the problem, however they are all based only on raw, unlabeled video frames and do not leverage the information of bounding boxes manually annotated by human raters.

3. Approach

For stages 1 and 2 (object detection and object tracking) we used the publicly available implementation of YOLOv5 + StrongSORT and OSNet [11]. We found that adjusting the training parameters and the train-validation split of the dataset are critical for the final results of the model.

3.1. Train-validation split of the dataset

The task organizers provided a sample code for object detection using YOLOv5 on the task dataset. In the solution, the full dataset was split by taking 16 full videos as a training set and the reminding 4 videos as a validation set. We found that in this setup the network learns well the training set, while for the validation the precision stays at the level of $mAP_{0.5} \sim 14\%$, $mAP_{0.5:0.95} \sim 4\%$. This means that the differences between the videos from the training set and the validation set are large and the network does not generalize well beyond the training set.

The other split we tried was to take every 5th frame (frame 0, 5, 10, 15, ...) as a validation set while keeping the rest as a training set. In this setup the network quickly overfits and reaches the precision of $mAP_{0.5} \sim 99.5\%$, $mAP_{0.5:0.95} \sim 90\%$ on the validation set. This means that the validation set is too similar to the training set and we are not able to verify how much the network generalizes beyond the training set and hence we cannot optimize the training parameters.

Finally, we decided to construct the training set from the first 24 s of each video and the validation set from the last 6 s of each video. In this setup the network achieved the accuracy of $mAP_{0.5} \sim 91\%$, $mAP_{0.5:0.95} \sim 66\%$ on the validation set. In this approach we are sure that the network has a chance to learn on each of the different videos and at the same time the validation set is independent enough that we can adjust the training parameters and analyze their impact on the precision on the validation set.

3.2. Confidence threshold for object detection

We noticed that for the default confidence threshold (`conf_thres`) equal to 0.25 for object detection with YOLOv5, the trained model detects more sperm per frame that it should, according to a manual assessment — it returns on average 78 sperm per frame in the test set, while there is only 22 sperm per frame in the labeled training set. We have manually estimated the actual number of sperm in the test videos and compered it to the average number of sperm returned by the model for a few different thresholds and found out the most accurate number of sperm per frame are returned for threshold of 0.75. The average number of sperm detected for different thresholds are presented in table 1.

3.3. Computing the motility of sperm

To compute the motility of sperm, we use a novel approach of comparing the average (vector) velocities of cells for two different sampling intervals: T_{long} and T_{short}. The approach is based on the observation, that progressive sperm have high velocity, independent of sampling interval,

`conf_thres`	avg. number of cells in the test videos
0.25	78
0.50	67
0.70	44
0.75	**31**
0.80	17

Table 1
Average number of sperm detected per frame in the test videos for different confidence thresholds.

while non progressive sperm have high velocity when sampling on short interval but low velocity when sampling on longer intervals. We determine the motility of a sperm by comparing its two average velocities v_{long}, v_{short} and comparing them with two threshold velocities: v_{imm} and v_{progr}. The exact formula is as follows:

$$motility = \begin{cases} immotile & \text{if} \quad v_{short} \leq v_{imm}, \\ nonprogressive & \text{if} \quad v_{short} > v_{imm} \wedge v_{long} \leq v_{progr}, \\ progressive & \text{if} \quad v_{short} > v_{imm} \wedge v_{long} > v_{progr}. \end{cases}$$

We found that the most accurate results are obtained for $T_{short} = 5/50\,\text{s}$, $T_{long} = 20/50\,\text{s}$, $v_{imm} = 0.003 v_{max}$ and $v_{progr} = 0.01 v_{max}$ where v_{max} is the maximum sperm velocity, measured at the 1/50 s sampling intervals.

3.4. Finding the fastest sperm

We achieved the best accuracy in determining the fastest sperm by comparing the average velocity computed at 4 s sampling intervals. Moreover, to filter out the outliers coming from wrongly tracked sperms (the same sperm id jumping between two different sperm cells), we skip the cells with standard deviation of the average velocity higher than 0.05.

4. Results and Analysis

For sub-task 1, the performance of the model was evaluated using the HOTA metric [12]. This metric combines two components: *Localization Accuracy* and *Association Accuracy* and therefore can serve a single number to quantify the performance of both, detection and tracking parts of the workflow. The detailed HOTA numbers for detecting normal sperm are presented in table 2.

seq	$HOTA_{AUC}$
66	23.8%
68	24.6%
73	29.0%
76	37.6%
80	28.6%
combined	28.3%

Table 2
$HOTA_{AUC}$ metrics for detecting normal sperm in the test set of 5 videos.

Sub-task 2 was evaluated by measuring different statistical errors between the predicted and ground-truth distribution of progressive/non-progressive/immotile sperm. The results are presented in table 3.

variable	mean absolute error	mean squared error	root mean squared error	root squared log error	median absolute error
progressive %	15.6	295	17.2	1.36	17
non progressive %	8.6	84	9.2	0.26	10
immotile %	22.2	551	23.5	0.24	23
average %	15.5	310	16.6	0.62	17

Table 3
Errors in predicting the progressive/non-progressive/immotile distribution.

References

[1] V. Thambawita, S. Hicks, A. M. Storås, J. M. Andersen, O. Witczak, T. B. Haugen, H. Hammer, T. Nguyen, P. Halvorsen, M. A. Riegler, Medico Multimedia Task at MediaEval 2022: Transparent Tracking of Spermatozoa, in: Proceedings of MediaEval 2022 CEUR Workshop, 2022.

[2] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, CoRR abs/1506.02640 (2015). URL: http://arxiv.org/abs/1506.02640. arXiv:1506.02640.

[3] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, CoRR abs/1703.07402 (2017). URL: http://arxiv.org/abs/1703.07402. arXiv:1703.07402.

[4] Y. Du, Y. Song, B. Yang, Y. Zhao, Strongsort: Make deepsort great again, 2022. URL: https://arxiv.org/abs/2202.13514. doi:10.48550/ARXIV.2202.13514.

[5] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, CoRR abs/1905.00953 (2019). URL: http://arxiv.org/abs/1905.00953. arXiv:1905.00953.

[6] J. Cao, X. Weng, R. Khirodkar, J. Pang, K. Kitani, Observation-centric sort: Rethinking sort for robust multi-object tracking, 2022. URL: https://arxiv.org/abs/2203.14360. doi:10.48550/ARXIV.2203.14360.

[7] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, X. Wang, Bytetrack: Multi-object tracking by associating every detection box, CoRR abs/2110.06864 (2021). URL: https://arxiv.org/abs/2110.06864. arXiv:2110.06864.

[8] S. Hicks, J. Andersen, O. Witczak, V. Thambawita, H. Hammer, T. Haugen, M. Riegler, Machine learning-based analysis of sperm videos and participant data for male fertility prediction, Scientific Reports (2019). doi:10.1038/s41598-019-53217-y.

[9] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, T. B. Haugen, Extracting temporal features into a spatial domain using autoencoders for sperm video analysis, CoRR abs/1911.03100 (2019). URL: http://arxiv.org/abs/1911.03100. arXiv:1911.03100.

[10] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, T. B. Haugen, Stacked dense optical flows and dropout layers to predict sperm motility and morphology, 2019. URL: https://arxiv.org/abs/1911.03086. doi:10.48550/ARXIV.1911.03086.

[11] M. Broström, Real-time multi-camera multi-object tracker using yolov5 and strongsort with osnet, https://github.com/mikel-brostrom/Yolov5_StrongSORT_OSNet, 2022.

[12] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, B. Leibe, Hota: A higher order metric for evaluating multi-object tracking, International Journal of Computer Vision (2020) 1–31.