Medico Multimedia Task at MediaEval 2022: Transparent Tracking of Spermatozoa

Vajira Thambawita[1], Steven Hicks[1,2], Andrea M Storås[1,2], Jorunn M Andersen[2], Oliwia Witczak[2], Trine B Haugen[2], Hugo Hammer[1,2], Thu Nguyen[1], Pål Halvorsen[1,2] and Michael A Riegler[1,3]

[1]*SimulaMet, Norway*
[2]*OsloMet, Norway*
[3]*UiT The Arctic University of Norway, Norway*

Abstract

The *Medico: Multimedia for Medicine Task* is running for the sixth time as part of MediaEval 2022. This year, the task focuses on automatically detecting and tracking spermatozoa in video recordings of human semen. We provide a dataset with 20 video recordings of spermatozoa with manually annotated bounding-box coordinates and a set of sperm characteristics. The task consists of four subtasks.

1. Introduction

The 2022 Medico task tackles the challenge of tracking spermatozoa cells in video recordings of untreated semen samples. Manual evaluation of sperm motility using a microscope is time-consuming and expensive, and requires experts who have extensive training. Furthermore, the validity of manual sperm analysis becomes unreliable due to limited reproducibility and high inter-personnel variations due to the complexity of tracking, identifying, and counting sperm in fresh samples. The existing computer-aided sperm analyzer systems are not working well enough for application in a real clinical setting due to unreliability caused by the consistency of the semen sample. Therefore, we need to research new methods for automated sperm analysis.

The goal is to encourage task participants to track individual spermatozoa in real-time and combine different data sources to predict common measurements used for sperm quality assessment, specifically the motility of the spermatozoa.

We hope that this task will encourage the multimedia community to aid in the development of computer-assisted reproductive health and discover new and clever ways of analyzing multimodal datasets. In addition to good analysis performance, an important aspect is also the efficiency of the algorithms due to the fact that the assessment of the sperm is performed in real-time and therefore requires real-time feedback. Furthermore, adding explainability to predictions of machine learning algorithms can improve the trust between medical experts and machine learning solutions.

For the task, we will provide a dataset that contains in total 20 30-second videos, a set of sperm characteristics (hormones, fatty acids data, etc.), frame-by-frame bounding box annotations, some anonymized study participants-related data, and motility and morphology data following the WHO recommendations.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online
vajira@simula.no (V. Thambawita); steven@simula.no (S. Hicks); andrea@simula.no (A. M. Storås); joran@oslomet.no (J. M. Andersen); oliwiaw@oslomet.no (O. Witczak); tribha@oslomet.no (T. B. Haugen); hugoh@oslomet.no (H. Hammer); thu@simula.no (T. Nguyen); paalh@simula.no (P. Halvorsen); michael@simula.no (M. A. Riegler)

Based on this data, the participants will be asked to solve the following four subtasks, where subtasks 3 and 4 are optional:

- **Subtask 1:** The goal of this subtask is real-time tracking of sperm cells in a given semen video. Tracking should be performed by predicting bounding box coordinates with the similar format to the bounding box coordinates provided with the development datasets. In this task, models should track sperm in each frame of a provided video in real-time. Therefore, frames per second is an important factor to measure.
- **Subtask 2:** The goal of this task is to predict sperm motility[1] in terms of the percentage of progressive and non-progressive spermatozoa. The prediction needs to be performed patient wise, resulting in one value per patient per predicted attribute. Sperm tracking or bounding boxes predicted in task 1 are required in order to solve this task.
- **Subtask 3:** This task focuses on identifying the fastest sperm cells with corresponding average speed and highest top speed. One specific challenge with this subtask is that the video also changes the view of the sample. This happens because the sample is moved below the microscope to observe the complete sample area. Therefore, the tracking has to be performed per viewpoint on the sample.
- **Subtask 4:** This subtask builds on task 1, 2 and 3. The goal is to provide an explanation of the model predictions to convince domain experts about the correctness of the results. There are not any specific requirements for this task, and participants can choose for which subtask they want to provide the explanations. A report should be provided that can be given to medical domain experts.

For subtasks 2 and 3, the participants are encouraged to consider the temporal aspect of the videos. This is important due to the fact that single-frame-based analysis will not be able to catch the movement of the sperm (motility), which contains important information to perform proper predictions on subtasks 2 and 3.

2. Dataset Details

The 2022 Medico task uses the data set VISEM [1], which contains data from 85 male patients aged 18 years or older. For this task, we provide annotations for 30 seconds video clips from selected 20 videos and we call this new dataset as VISEM-Tracking [2].

For each patient, we include a video of live sperm (video and extracted frames), manually annotated bounding box details for each spermatozoon, a set of measurements from a standard semen analysis for the whole sample, a sperm fatty acid profile, the fatty acid composition of serum phospholipids, study participants-related data, and WHO analysis data. The bounding box coordinates are provided in two separate folders, one has bounding box coordinates in YOLO format [3] and the other folder contains feature identifiers in addition to the bounding box coordinates. These feature identifiers can be used to identify the same bounding box in different frames in a video. Each video has a resolution of 640x480 and runs at 50 frames per second. The dataset contains five CSV files and folders containing the videos and bounding box data. The name of each video file contains the video's ID, the date it was recorded, and a small optional description. Then, the end of the filename contains the ID of the person who assessed the video. Furthermore, VISEM contains five CSV files for each of the other data provided, a CSV file with the IDs linked to each video, and a text file containing * descriptions of some of

[1]Motility is the ability to move independently, where a progressive spermatozoon is able to "move forward" and a non-progressive would move for example in circles without any forward progression.

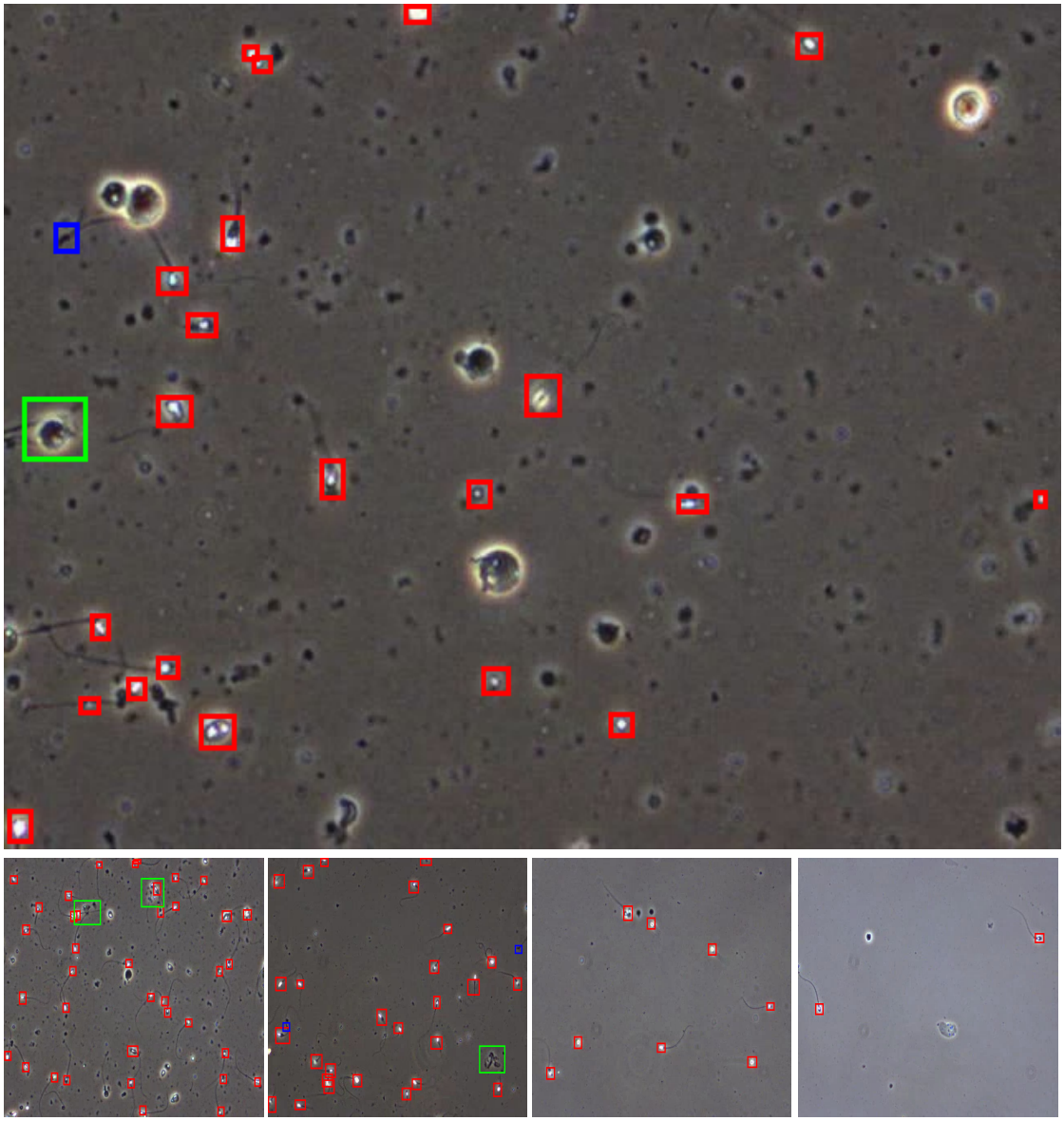

Figure 1: Sample video frames with corresponding bounding boxes. Top: big image showing different classes of bounding boxes, red - sperm, green - sperm cluster and blue - small or pinhead sperm. Bottom: presenting different sperm density levels from high to low (from left to right, respectively).

the columns of the CSV files. One row in each CSV file represents a participant. The provided CSV files are:

- semen_analysis_data: The results of standard semen analysis.
- fatty_acids_spermatozoa: The levels of several fatty acids in the sperm of the participants.
- fatty_acids_serum: The serum levels of the fatty acids of the phospholipids (measured from the blood of the participant).
- sex_hormones: The serum levels of sex hormones measured in the blood of the participants.
- study_participant_related_data: General information about the participants such as age, abstinence time, and Body Mass Index (BMI).

All study participants agreed to donate their data for the purpose of science and provided the necessary consent for us to be able to distribute the data (checked and approved by the Norwegian data authority and ethical committee).

3. Evaluation

For the first sub-task, sperm cell tracking, the mAP (mean average precision) will be calculated as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{1}$$

with two different threshold values of 0.5 and $0.5 : 0.95$. For the second range of threshold values, the average of incremental threshold values between 0.5 and 0.95 by 0.05 will be calculated. Finally, a fitness value will be calculated as follows:

$$fitness\_value = 0.1 \times mAP_{0.5} + 0.9 \times mAP_{0.5:0.95}. \tag{2}$$

In addition to the mAP and fitness values, we will calculate frames per second (FPS) for the first task to evaluate the speed of predictions. For the evaluation of the second task, we calculate the mean squared error and mean absolute error. Additionally, the time required to predict the output of the subtask 2 is also calculated.

The optional third and fourth tasks will be evaluated using manual evaluation with the help of three different experts within human reproduction. A fixed hardware setup will be used for all evaluations to measure the speed of predictions.

4. Discussion and Outlook

Previous studies of applying deep learning solutions [4, 5, 6] in sperm analysis show the potential of applying deep learning techniques to predict morphology and motility levels of given sperm videos. However, the high values of mean average errors of the studies imply that pure video input to deep learning models makes predictions less accurate. In this regard, we provide additional manually annotated bounding boxes for three classes, namely sperm, sperm clusters and small or pinhead sperms. We believe that these additional localization details will help to improve the deep-learning-based sperm analysis, which again will help future assisted human reproduction.

References

[1] T. B. Haugen, S. A. Hicks, J. M. Andersen, O. Witczak, H. L. Hammer, R. Borgli, P. Halvorsen, M. A. Riegler, Visem: A multimodal video dataset of human spermatozoa, in: Proceedings of the ACM on Multimedia Systems Conference (MMSys), 2019. doi:10.1145/3304109.3325814.

[2] V. Thambawita, S. A. Hicks, A. M. Storås, T. Nguyen, J. M. Andersen, O. Witczak, T. B. Haugen, H. L. Hammer, P. Halvorsen, M. A. Riegler, Visem-tracking: Human spermatozoa tracking dataset, arXiv preprint arXiv:2212.02842 (2022).

[3] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788. doi:10.1109/CVPR.2016.91.

[4] S. A. Hicks, J. M. Andersen, O. Witczak, V. Thambawita, P. Halvorsen, H. L. Hammer, T. B. Haugen, M. A. Riegler, Machine learning-based analysis of sperm videos and participant data for male fertility prediction, Scientific reports 9 (2019) 1–10.
[5] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, T. B. Haugen, Extracting temporal features into a spatial domain using autoencoders for sperm video analysis, arXiv preprint arXiv:1911.03100 (2019).
[6] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, T. B. Haugen, Stacked dense optical flows and dropout layers to predict sperm motility and morphology, arXiv preprint arXiv:1911.03086 (2019).