UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD thesis

Malte Stær Nissen — nissen@di.ku.dk

Automatic Motility Analysis of Human Sperm

Supervisors: Mads Nielsen, Christian Igel, Kristian Almstrup, Søren Kjærulff, and Torben Trindkær Nielsen

August 1, 2018

Contents

Preface

This thesis is a Public Sector Industrial PhD submitted for the degree of Doctor of Philosophy at the Faculty of Science, University of Copenhagen, Copenhagen, Denmark. The thesis was conducted as a collaboration between the three parties:

- Department of Computer Science, University of Copenhagen, Copenhagen, Denmark (DIKU),
- Department of Growth and Reproduction, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark (GR),
- ChemoMetec A/S, Allerød, Denmark (CM).

Being a Public Sector Industrial PhD, the project was aimed at research within the needs of the industrial partners GR and CM.

GR is a department at the public hospital Rigshospitalet conducting clinical workups of patients and research within male reproduction. Their routine semen quality analyses are primarily conducted manually and are very labour intensive.

CM is a private company producing image cytometers for automatic counting of cells. In general, their image cytometers take microscopy images of biological samples, identify cells by segmentation, and present scatter plots of various cell parameters such as cell area, perimeter, and mean intensity. CM developed their Xcyto® 10[1] (XC10) image cytometer while these studies were conducted.

We based the project on GR's need for automating semen quality analyses and CM's need for developing image and video analysis for the XC10. A key requirement was to use the XC10 image cytometer (and developmental prototypes hereof) for acquisition of data (microscopy images/video). Therefore any cell detection algorithm developed in the project had to include an estimation of a segmentation mask for each detected object in order for the algorithm to be usable in the XC10 afterwards.

The studies were conducted at the locations of all three project partners listed above and at the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC - University Medical Center Rotterdam, Rotterdam, the Netherlands between October 2014 and July 2018

This PhD thesis is written as a synopsis containing the following three studies:

[1] `https://chemometec.com/automated-cell-counters/xcyto-10/`

- Malte S. Nissen, Oswin Krause, Kristian Almstrup, Søren Kjærulff, Torben T. Nielsen, and Mads Nielsen. "Convolutional Neural Networks for Segmentation and Object Detection of Human Semen". In: *Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12–14, 2017, Proceedings, Part I*. ed. by Puneet Sharma and Filippo Maria Bianchi. Cham: Springer International Publishing, 2017, pp. 397–406. ISBN: 978-3-319-59126-1. DOI: 10.1007/978-3-319-59126-1_33. URL: http://dx.doi.org/10.1007/978-3-319-59126-1_33

- Malte S. Nissen, Oswin Krause, Kristian Almstrup, Søren Kjærulff, Erik Meijering, and Mads Nielsen. "Estimation of motility distribution: A study of linearity for human sperm motility analysis". Journal paper to be submitted for Transactions on Medical Imaging (T-MI). 2018

- Malte S. Nissen, Mads Nielsen, and Kristian Almstrup. "Evaluation of a new integrated and fully automated system for sperm motility analysis". Journal paper to be submitted for Andrology. 2018

This work is partly funded by the Innovation Fund Denmark (IFD) under File No. 4135-00169B and partly funded by ChemoMetec A/S, Gydevang 43, DK-3450 Allerød, Denmark.

Acknowledgements

I would like to thank my main supervisor, Mads Nielsen, and my co-supervisor, Kristian Almstrup, for providing excellent supervision throughout the project and supporting me during the final preparations of my thesis. Many thanks to my industrial co-supervisor, Søren Kjærulff, for providing excellent supervision and insight into the world of image cytometry. An honorable mention goes to to Lars D. Johansen for helping with the practical use of the Xcyto 10 prototypes and final machine. I would like to thank Erik Meijering for great supervision and for hosting me during my stay abroad. A special and thank to Ihor Smal for all your help especially during my stay abroad.

I thank my colleagues in the (now re-structured) Image Group, at Rigshospitalet, and at ChemoMetec for being great colleagues. Especially, thanks to Raghavendra Selvan, Oswin Krause, Silas Ørting, Niklas Kasenburg, and Dídac Rodríguez Arbonès, and Akshay Pai for interesting discussions, helpful advice, and creating a great social atmosphere.

A special thanks to my family, Bente, Jörgen, Ane, and Jonatan for always being there for me, asking interesting questions along the way. A big thanks to all my friends who have made the last 3.5 years great and exciting. Many thanks to my roommate Daniele for great discussions, having coped with my complaints for the last four years, and in general for being an awesome friend and roommate.

The biggest of thanks to Ghazal Alavioon, who has supported and helped me through the last part of my studies. You are my pillar of strength.

Summary

In this thesis we investigated automatic motility analysis of human semen. The investigation was conducted in three studies.

First, we investigated how to detect and segment sperm cells in bright field microscopy images from the Xcyto® 10 image cytometer. We developed a pixel-wise segmentation and detection algorithm based on the use of convolutional neural networks achieving high pixel-wise accuracy, precision and recall.

Second, we studied how to conduct an unbiased estimation of the motility distribution of sperm cells and whether we can track sperm cells sufficiently reliably to obtain accurate motility distributions in practice. The study was conducted by analysing a set of semi-automatically annotated sperm cell tracks. Based on the study we recommended a set of guidelines for conducting unbiased motility estimation. We combined our detector from the first study with an existing linker method to obtain an automatic method for tracking of human sperm cells. Using this tracker we obtained motility distributions nearly identical to the theoretical distributions.

Third, we evaluated the automatic system for conducting motility analysis of human sperm by comparing it with manual motility analysis resulting in comparable results. However, more data needs to be collected before finally concluding whether the system can be used during routine analysis.

Resumé

I denne afhandling undersøgte vi automatisk motilitetsanalyse af menneskesæd. Undersøgelsen blev gennemført i tre studier.

Først undersøgte vi, hvordan man genkender og segmenterer sædceller i bright field mikroskopibilleder fra Xcyto® 10 billedcytometret. Vi udviklede en pixel-wise segmenterings- og detektionsalgoritme baseret på brugen af konveksionale neurale netværk, som opnåede høj pixel-mæssig nøjagtighed, precision og recall.

I det andet studie undersøgte vi hvordan man foretager en upartisk vurdering af motilitetsfordelingen af sædceller, og om vi kan spore sædceller tilstrækkeligt pålideligt til at opnå en nøjagtig motilitetsfordeling i praksis. Undersøgelsen blev udført ved at analysere et sæt semi-automatisk annoterede tracks a sædceller. Baseret på undersøgelsen anbefalede vi et sæt retningslinjer for gennemførelse af objektiv motilitetsestimation. Vi kombinerede vores detektor fra det første studie med en eksisterende linker metode for at opnå en automatisk metode til sporing af humane sædceller. Ved hjælp af denne tracker opnåede vi motilitetsfordeling næsten identiske med de teoretiske fordelinger.

I det tredje studie evaluerede vi det automatiske system til udførelse af motilitetsanalyse af humane sædceller ved at sammenligne den med manuelle motilitetsanalyse, hvilket resulterede i sammenlignelige resultater. Der skal imidlertid indsamles mere data for at komme med en endelig konklusion på, om systemet kan bruges i praksis.

Part I

Synopsis

Chapter 1

Introduction

Human fertility has become a major concern over the past decades. An estimate of 48.5 million couples (2010) are affected by infertility [27] leading to a large demand for fertility analysis for both males and females.

The fertility of the male is investigated by conducting a clinical semen quality analysis (SQA) of the ejaculate as defined by the WHO [51]. SQA consists of estimating semen quality parameters such as pH-value, viscosity, volume, spermatozoa (sperm) count (concentration and total count), and sperm quality, where sperm quality is quantified by the sperm motility, sperm morphology, and other traits specific to individual sperm cells. The fertilisation capabilities of the male is a complex combination of all these parameters ultimately summing up to whether or not the male is able to fertilise the oocyte (egg) of the female and get healthy offspring.

A meta-study of the temporal trend in sperm counts was recently published [23] revealing a significant decline in sperm counts from 1973 to 2011. Figure 1.2 shows the meta-study regression lines for mean sperm concentration (a) and mean total sperm count (b) as a function of year of sample collection presented by Levine et al. [23]. The decline was especially pronounced for the group of unselected western (men from North America, Europe, Australia, and New Zealand with no selection based on fertility) as seen in the graphs with declines of 52.4% and 59.3% for sperm concentration and total sperm count respectively. Furthermore, Skakkebaek et al. [45] recently highlighted the issues of male infertility.

A large decrease in fertility rate (average number of births per woman) has been observed since 1960 as seen in Figure 1.1. Multiple factors affect the fertility rate such as lifestyle, demographics, and ability to reproduce. Even if the ability to reproduce was not the cause of the drop in fertility rate, fertility still plays an increasingly important role. Fertility naturally plays a key role for couples trying to reproduce, but it plays an even more important role for society in order to be able to maintain the population in case fewer people choose to reproduce or they choose to get fever children.

The study of Levine et al. [23], the data from The World Bank, and the data from Dansk Fertilittsselskab alltogether highlight the importance of research within

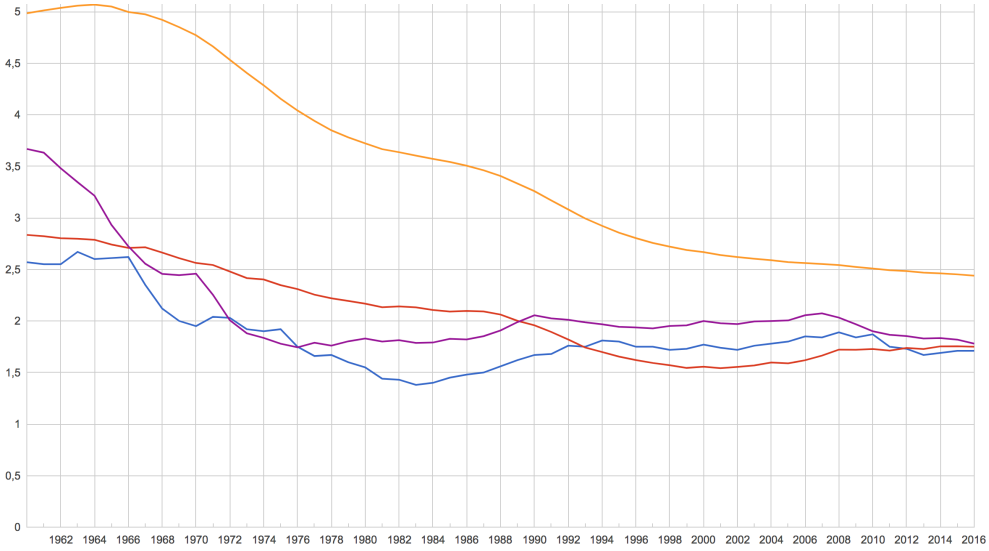


Figure 1.1: Fertility rate (average number of births per woman) between 1960 and 2016. Data is shown for the world (yellow), North America (purple), Europe and Central Asia (red), and Denmark (blue). Reproduced from Google Public Data with data from The World Bank.

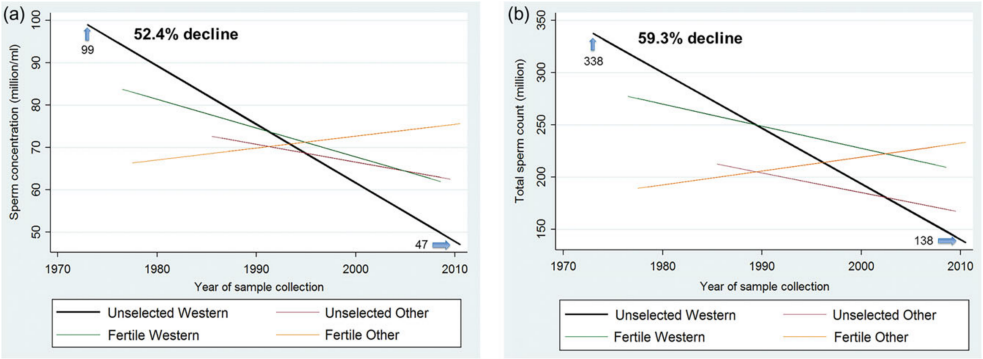


Figure 1.2: Meta-study regression lines for mean sperm concentration (a) and mean total sperm count (b) as function of year of sample collection. The figure has been reproduced from Levine et al. [23, Figure 3] by permission of Oxford University Press and the European Society of Human Reproduction and Embryology (ESHRE).

fertility and in particular male fertility.

As mentioned above, there are multiple important parameters affecting male fertility. Good sperm swimming abilities (progressive motility) is a proven predictor of fertility rate [33, 20]. While comparable historical sperm concentration data exists, issues with standardisation and reliability of motility analysis has caused an absence in international comparable historical motility data.

In these studies we focused on the sperm motility analysis requiring replicate counting of minimum 200 sperm cells [51]. The analysis is both labour intensive and

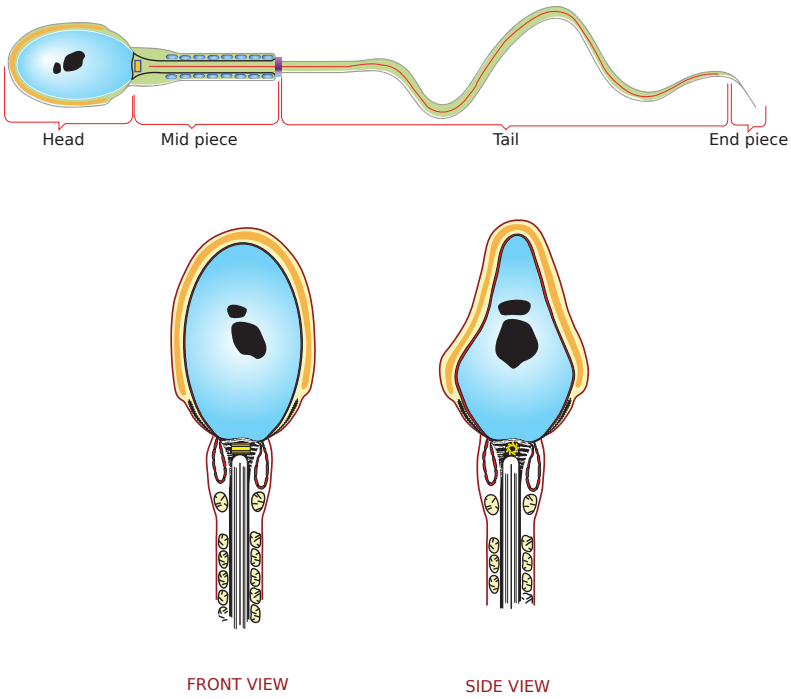


Figure 1.3: Illustration of human spermatozoa consisting of the head, mid piece, tail, and end piece (top). The head is illustrated both from a front (bottom left) and side view (bottom right). Original by Mariana Ruiz Villarreal spermatozoa [Public domain], via Wikimedia Commons.

prone to high inter-observer variation, though the latter is reducible to an acceptable range with proper quality control and training [13]. Furthermore the motility analysis needs to be conducted within an hour of ejaculation putting a strict bound on the number of samples each laboratory technician can process.

1.1 The sperm cell

Identification of sperm cells is the basis of motility analysis. Figure 1.3 shows an illustration of a sperm cell consisting of the head, mid piece, tail (flagellum), and end piece. The sperm cell head should be oval (front view) and normal sperm cells have been reported to have median length $l = 4.1\,\mu\text{m}$ (95% confidence interval 3.7-4.7) and median width $w = 2.8\,\mu\text{m}$ (95% confidence interval 2.5-3.2) [51, pp. 68, Section 2.15.1] or approx. $w = \frac{2}{3}l$. The mid piece should be approximately the same length as the head, and the tail should be approximately $45\,\mu\text{m}$ long.

Sperm cells swim using their tail for propulsion. The tail beat resembles propagation of a wave in 3D down through the tail. Given a 2D projection of the cell we mainly observe the motion as a 2D beating from side to side with 2-3 bends on the tail.

In order to

1.2 Computer-Aided Sperm Analysis

Research and development of computer-aided sperm analysis (CASA) for assisting laboratory technicians with SQA has been a topic of interest since the 1980s [26, 28, 25], with the Automated Semen Analyzer (CellSoft^{TM}, 1985) and HTM-2000 (Hamilton-Thorne) being the first commercially available CASA systems as described by Horst, Mortimer, and Mortimer [17]. Horst, Mortimer, and Mortimer [17] likewise described both biological and technical issues related to using CASA for human SQA such as dirty ejaculates with lots of debris causing mis-classification of non-sperm objects, high viscosity making representative sampling difficult, and difficult tracking of sperm cells through collisions causing inaccurate motility estimates. Therefore, no CASA currently exists which is globally accepted and adopted by human fertility clinics.

Several of the CASA-related issues have been studied with promising results lately such as the problem of predicting motility grading from sperm cell track kinematic metrics [12] and multi-object tracking of sperm cells with collision handling [12, 49].

Multiple systems have been developed and released during the last few years such as CEROS II and IVOS II with the Human Motility II software (Hamilton Thorne), QualiSperm® (Biophos), Sperm Class Analyzer® (Microptic), and SQA-Vision (Medical Electronic Systems), all of which show improvements over previous iterations of CASA systems.

Machine learning and computer vision has experienced a general improvement in performance during the past years due to the success of deep learning [21]. These methods helped improve state-of-the-art performance greatly on problems such as general image classification [19, 44, 48, 15], traffic sign classification/recognition [47], and mitosis detection [4]. We hoped to continue these advances by studying and applying deep learning techniques for solving issues within SQA.

Recently, Palme et al. [34] validated the use of image cytometry and DNA staining for sperm concentration estimation. We built upon recent successes by combining machine learning and computer vision with image cytometry for solving problems within automation of SQA. We did so by using the XC10 image cytometry for automatic acquisition of images and video of human semen samples, which we used as a basis for the SQA studies presented in this thesis. Figure 1.4 shows an image of the completely integrated XC10 used for image and video acquisition.

The thesis is structured as follows: First we briefly state the aim of the thesis in Chapter 2. Second, we describe the three datasets developed for the thesis in Chapter 3. Third, we go through the methods related to the three studies in Chapter 4. Fourth, we summarise the three studies conducted in Chapter 5. Fifth, we discuss the results of the studies in Chapter 6, and finally, we conclude upon the thesis in Chapter 7.

Figure 1.4: The XC10 image cytometer

Chapter 2

Aim

The aim of this thesis is to investigate and solve problems related to automating SQA analyses using new advances within machine learning and computer vision while taking into account the needs of CM. We investigated three questions with this aim in mind:

1. How can we detect and segment human sperm cells in a bright field microscopy image from the XC10 cytometer in less than a second using deep neural networks?

2. How can we compute an unbiased estimate of the sperm cell population motility and can we track sperm cells sufficiently reliably to obtain accurate motility distributions?

3. How does the automatic motility analysis compare to the manual analysis, and how precise is it?

These questions are addressed in the three studies briefly described in Chapter 5 and shown in full in Part II.

Chapter 3

Datasets

As part of our work we developed three datasets: the segmentation dataset, the tracking dataset, and the clinical motility dataset. In this chapter we briefly describe the microscope setup followed by a description of the data collection and ground truth annotations of the datasets. Finally, we briefly describe the ethical considerations related to data collection and handling.

3.1 Microscope setup

We used the XC10 image cytometer for capturing images and videos with bright field light setup and $20\times$ optical magnification.

Bright field microscopy makes visual identification of the sperm cells relatively easy by maintaining the overall visual features intact as long as the sperm cell is close to being in focus. Using regular light microscopy also means affecting the sperm cells minimally making motility analysis possible. The down-side of using bright field is, that the sperm cells change appearance drastically depending on their focus plane offset. Detection algorithms therefore have to take the change of appearance into account. The choice was influenced by an original goal of wanting to conduct morphology (appearance) analysis requiring a good representation of the full sperm cell. We later abandoned this goal due to an insufficient level of optical magnification making it difficult to meet the strict standards of morphology analysis described by the WHO [51, pp. 67-70, Section 2.15].

$20\times$ optical magnification yields a good compromise between level of sperm cell detail and field of view size. The combination of optical magnification and camera used in the XC10 gives a resolution of $0.2269\,\mu\mathrm{m/pixels}$ and a field of view of $435.6 \times 326.7\,\mu\mathrm{m}$. This makes it possible to distinguish sperm cells from debris while having > 650 sperm cells per field of view for the most concentrated samples.

The camera used by the XC10 is able to capture raw grayscale images at 14-bit and video at 8-bit depth. We therefore capture images at 8-bit and normalise them to cover the full 8-bit range in order to be able to use the same detection algorithm

for both still images and videos. The normalisation gives us images with grey background and clearly visible cells.

3.2 Segmentation dataset

We developed the segmentation dataset for being able to identify and segment sperm cells in bright field microscopy images.

Data collection

We collected a total of 35 individually independent samples in two data collection iterations. In the first iteration we collected four samples during December 2014. In the second iteration we collected 31 samples. between December 2015 and May 2016. Samples were pools of up to 4 anonymous ejaculates each. Each semen sample was collected by ejaculating into a plastic container and allowed to liquefy in an incubator at $37\,^{\circ}\mathrm{C}$ before approx $200\,\mu\mathrm{L}$ of each sample was extracted by pipette and placed in an Eppendorf tube. The samples were mixed thoroughly after being placed in an Eppendorf tube. The semen samples were cooled to approx $5\,^{\circ}\mathrm{C}$, and transported to CM in an insulated cooling container with a cooling bag for processing within 3 days. Samples were destructed immediately after processing.

We mixed a dilution medium of Bicarbonate-Formalin as described by World Health Organisation (WHO, et al. [51, Section 2.7.5, pp. 51] and added $10\,\mu\mathrm{g/mL}$ Hoechst-33342 (H342, ChemoMetec) dye to the dilution medium for staining of DNA as described by Egeberg Palme et al. [8]. We diluted each sample with the dilution medium to fixate the cells in the sample and to achieve views containing 2-290 sperm cells.

$30\,\mu\mathrm{L}$ of the sample was loaded into a $100\,\mu\mathrm{m}$ fixed-depth glass chamber and the cells were allowed to sediment to the bottom of the chamber. The dilution, fixation and sedimentation was done to achieve a suitable amount of non-overlapping sperm cells in each view and to force the sperm cells into the same focus plane. We then captured multiple spatial views of the sample at different spatial locations and for each of these views captured a z-stack of images around the focus plane estimated manually. For each bright field image captured, we also captured a fluorescence image of the Hoechst dye absorbed by the DNA in sperm cells or other cell types to aid the annotation process. A $405\,\mathrm{nm}$ LED was used to excite the Hoechst dye and the emitted light was collected using a band pass filter. In the initial iteration we only saved the z-stack image closest to the focus plane whereas we saved 3-4 images centred at the focus plane in the second iteration. Using images around the focus plane allowed us to capture the full range of variation in sperm cell appearance due to the bright field microscopy choice.

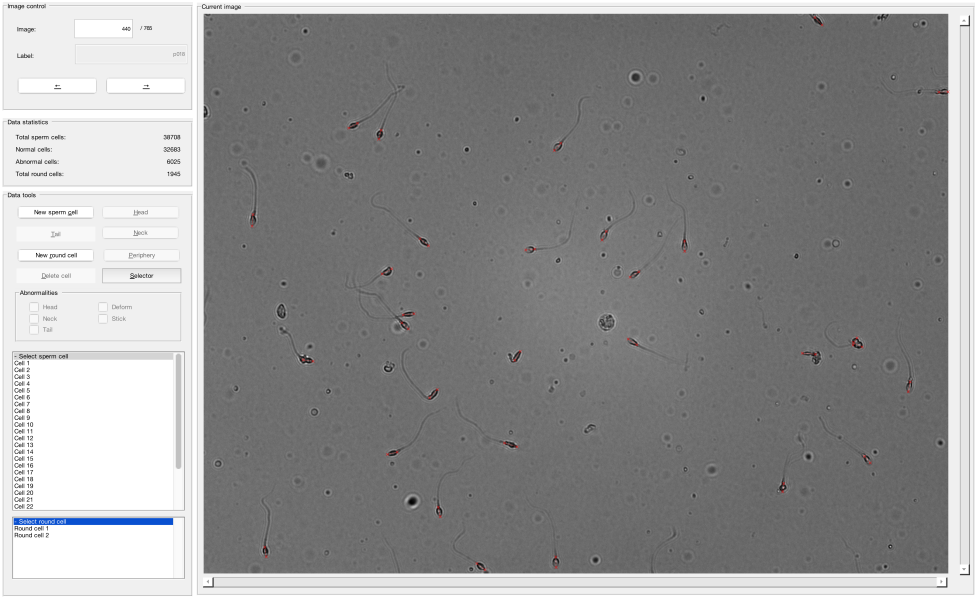


Figure 3.1: Segmentation GUI. The GUI consists of a window showing the image being annotated, controls to switch between images, statistics for the dataset, annotation tools, and a list of annotated sperm and round cells.

Ground truth annotation

The ground truth annotation was conducted in the same two iterations as the capturing of the images. The images of the dataset contain both sperm cells and various types of debris. The sperm cells are the most important cells in the samples, but round cells containing DNA (typically larger than sperm cells) are likewise interesting and make up a large portion of the visual variation in the samples. We therefore annotated both sperm and round cells in all images.

Sperm cells were annotated by their neck point and tip of their head, and round cells were annotated by an outline of their periphery. In the initial iteration we annotated the centre line of the sperm cell tails as well as simple morphological abnormalities of the sperm cells. After the initial iteration we decided to focus strictly on segmentation of the sperm cell heads and therefore only annotated the sperm cell head and round cell periphery. This choice resulted in a quicker annotation process making us able to annotate more samples.

Figure 3.1 shows a screenshot of the GUI we developed for manual annotation of sperm cells. We used this annotation tool to annotate the entire segmentation dataset. The images of the initial iteration (4 samples, 107 images) were annotated one by one. In the second iteration we first annotated the image mostly in focus from each z-stack, copied the annotations to the non-focussed images of each z-stack and corrected these ground truths where needed.

The result was a dataset of 765 fully annotated images containing a total of 38,708 sperm cells and 1,945 round cells. Table 3.1 shows an overview of semen

sample-level information about the segmentation dataset.

From the annotations we created automatically generated segmentation masks to be used as ground truths. In total we had three classes in our masks: background, sperm cell head, and round cell. Sperm cell head segmentations were created by exploiting our knowledge of their morphology. An ellipsoid with length l and width $2/3l$ was drawn for each sperm cell. The length was determined by the distance between the tip and neck points, which were also the end points of the ellipsoid. Round cell segmentations were created by filling in the area surrounded by each round cell periphery annotation. An example image can be seen in Figure 3.2 along with the corresponding Hoechst image, annotations and automatically generated segmentation mask. More examples of images and information regarding the segmentation dataset can be found in the article in Chapter 8.

3.3 Tracking dataset

We created the tracking dataset to develop and study tracking algorithms of sperm cells with the goal of achieving an accurate and unbiased semen sample motility estimation.

Data collection

We captured a total of 244 videos of 24 individually independent samples. 21 of these samples were pooled (from 3 individuals each) and three of the samples were individual donor samples. The samples were collected between December 2016 and January 2017 and collection followed the procedure of the segmentation dataset: ejaculation into a plastic container followed by liquefaction on a tilting table. Approx. $200\,\mu\mathrm{L}$ of each sample was transferred to an Eppendorf tube and transported to CM in an insulated box containing passive heating elements pre-heated to $37\,^{\circ}\mathrm{C}$. The samples reached CM where they were placed in an incubator at $37\,^{\circ}\mathrm{C}$ and processed for video capturing within four hours of ejaculation.

Processing consisted of either diluting the sample, conducting a swim-up procedure, or using the raw semen. Dilution was conducted by diluting an aliquot of the raw semen with Phosphate-buffered saline (PBS) and mixing the sample gently using a pipette. Five samples having an average number of automatically detected sperm cells per view of more than 200 were diluted to achieve less than 200 sperm cells per view. The dilution procedure was conducted to achieve samples with less than 120 sperm cells per view making them suitable for multi target tracking.

The long time since ejaculation caused a decrease in motility and an absence of highly motile semen samples. Therefore, a swim-up procedure of donor two samples was conducted to purify the semen resulting in samples with highly motile sperm cells.

The swim-up procedure was conducted by following Rehfeld, Dissing, and Skakkebæk [37]: “Motile spermatozoa were recovered from raw ejaculates by swim-up separation in human tubular fluid (HTF^{+}) medium containing 97.8 mM NaCl, 4.69 mM

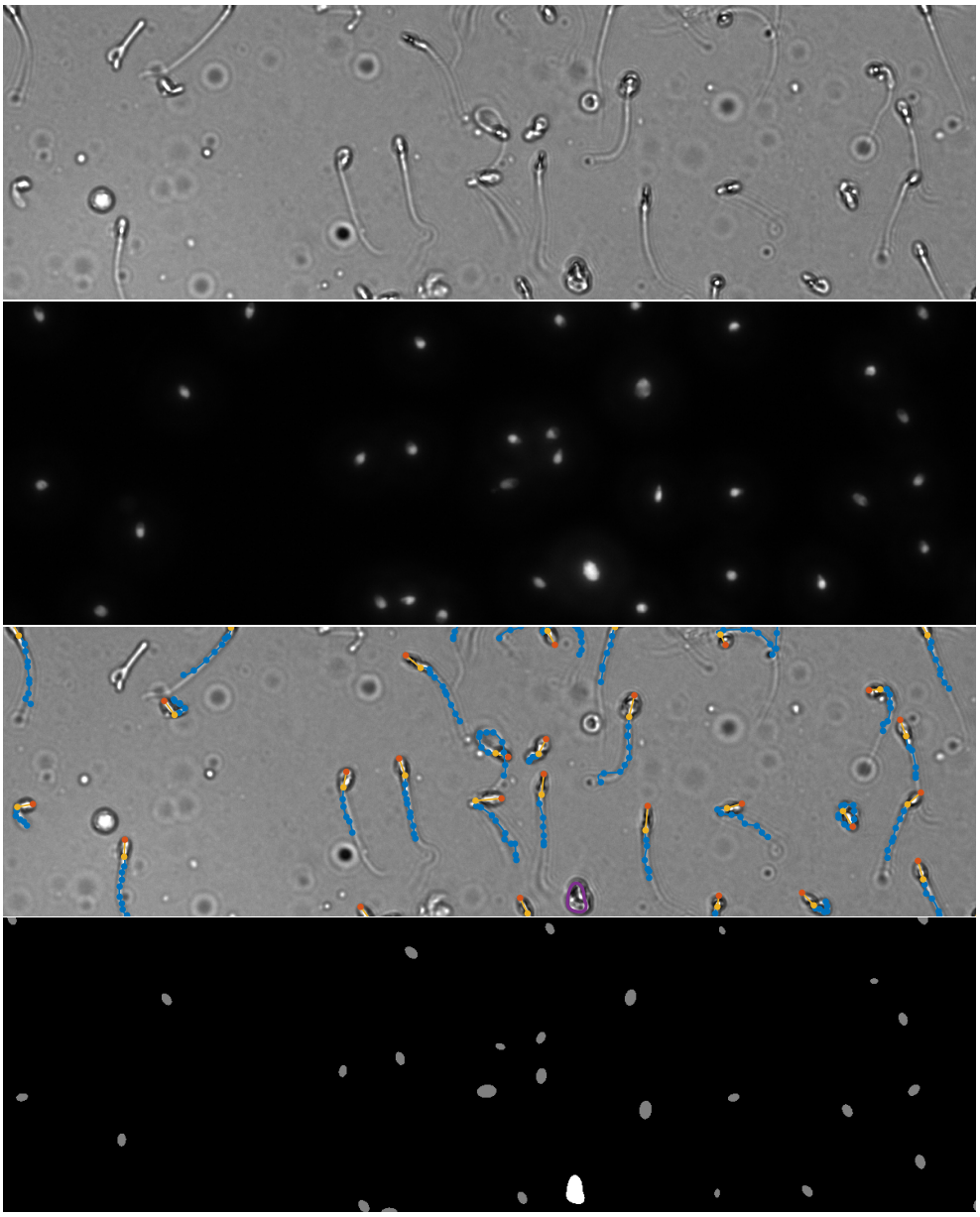


Figure 3.2: Cutout of example image from the segmentation dataset. The images shown on order of appearance: bright field image, hoechst image, annotations, and generated segmentation mask. For each sperm cell the tip of the head (red dot), neck point (orange dot), and tail (blue dots) were annotated. For each round cell the periphery (purple) was annotated. The segmentation mask shown contain the three classes: background (black), sperm cell head (grey), and round cells (white).

KCl, 0.2 mM $\mathrm{MgSO_4}$, 0.37 mM $\mathrm{KH_2PO_4}$, 2.04 mM $\mathrm{CaCl_2}$, 0.33 mM $\mathrm{Na-pyruvate}$, 21.4 mM $\mathrm{Na-lactate}$, 2.78 mM glucose, 21 mM HEPES, and 4 mM $\mathrm{NaHCO_3}$, adjusted to pH 7.3–7.4 with NaOH as described elsewhere [42]. After 1 hour at 37°C, the swim-up fraction was removed carefully and sperm concentration was determined by image cytometry as described in [9].". The sperm concentration determination step was skipped.

We conducted the following procedure when capturing videos of the samples:

1. Take an aliquot of the semen sample
2. Process the sample (dilution, swim-up, or raw semen)
3. Load processed aliquot into $20\,\mu\mathrm{m}$ glass slide and insert glass slide into XC10
4. Capture 8-10 views by conducting the following steps:
    a) Move to view position
    b) Conduct semi-automatic focusing
    c) Capture a video of 512 frames (20.48 seconds)

Ground truth annotation

We annotated individual sperm cell tracks in a subset of the videos captured. For each video we conducted the following steps:

1. Automatically detect sperm cells using the method of Nissen et al. [31]
2. Perform nearest neighbour tracking
3. Manually fix all "deaths" and "births" of tracks (typically at collision points)
4. Manually playback and fix errors using graphical user interface (GUI)

Before beginning the annotation process we developed a GUI for manually inspecting and correcting sperm cell tracks. The GUI supported import of the tracks and detections created by the first two steps of our semi-automatic annotation approach described above. The GUI is shown in Figure 3.3. We developed the GUI to speed up the annotation process by having buttons for focusing on short tracks, collision points, non-edge births of tracks, and frames in tracks without positions detected as well as key-bindings for all essential buttons.

At first we annotated one view (d01, 3x dilution) for full duration (512 frames) and one view for half duration (256 frames) taking approximately 8 and 4 hours respectively. The two samples had an average of 41.9 and 47.8 sperm cells per view respectively. More concentrated samples would take longer to annotate, so we decided to annotate 5 seconds (125 frames) in order to allow for a higher amount of samples and views to be annotated. Annotation of a single sample view for 125

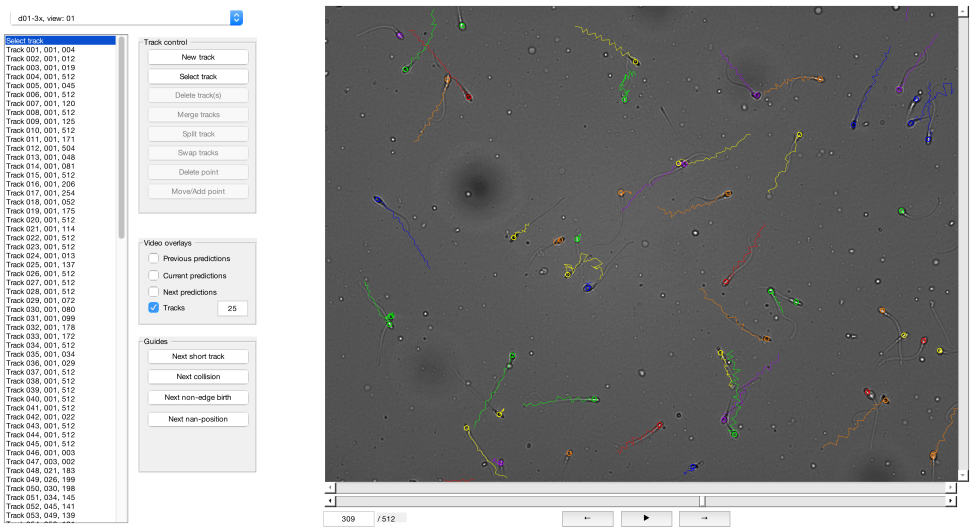


Figure 3.3: Tracking GUI. The GUI consists of a video player/frame window, a set of controls for annotating sperm cell tracks and a list of annotated tracks.

frames took between 2 and 8 hours. Towards the end of the annotation process we needed more full duration data for our analysis, and thus we annotated two additional samples (p02 and p09) for 512 frames (37 hours of annotation, each)

Table 3.2 shows an overview of information about the tracking dataset. Notice the average number of detected sperm cells per view ranges from 20.7 to 596.4, and the annotated videos contain between 21.5 and 122.0 sperm cells per view.

Figure 3.4 shows a visualisation of example image from a video in the tracking dataset with the annotated tracks for the last 25 seconds. More information regarding the tracking dataset can be found in the manuscript in Chapter 9.

3.4 Clinical motility dataset

The clinical motility dataset is the third and the last dataset we developed during the project. We created the dataset to clinically validate the automatic motility analysis by comparing it with manual motility analysis. The data collection is described in details in the manuscript in Chapter 10 and I therefore refrain from repeating it here.

3.5 Ethical considerations

All measurements were performed on excess from routine semen analysis from patients and from quality control donors. The purpose of the data collection was to evaluate and improve the routine semen analysis and the experiments were therefore considered a part of the routine analysis. All data was kept either fully anonymised or pseudo-anonymised. The link between personal information and pseudo-anonymised

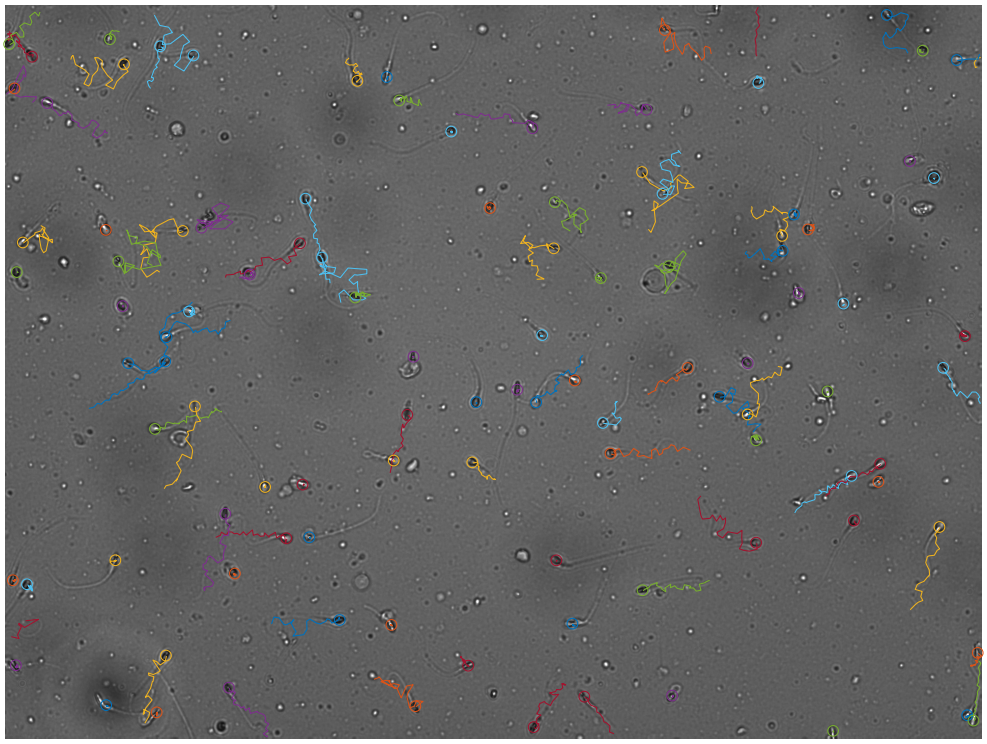


Figure 3.4: Example image from the tracking dataset with annotated sperm cell head positions (circles) and their tracks for the last 25 frames (lines) overlaid.

data was kept in a secure and access-restricted network drive at GR. Pseudo-anonymised samples were only processed at GR.

Label	# images	Patches ext.	# sperm	# round	Mean length	Std. length
1	47	118,428	4,721	94	24.93	4.77
2	20	44,844	199	90	20.69	6.63
3	20	56,603	2,963	47	23.20	4.18
4	20	54,645	886	34	26.06	3.90
p001	13	31,448	298	69	32.47	6.67
p002	9	24,000	252	12	24.36	3.84
p003	9	23,562	303	6	23.97	4.19
p004	11	33,000	582	47	26.52	6.58
p005	12	35,140	687	40	24.41	4.81
p006	13	32,000	227	12	25.51	5.44
p007	12	36,000	274	104	22.24	5.58
p008	12	34,566	204	105	24.18	5.16
p009	24	72,000	488	89	25.40	4.51
p010	24	72,000	615	150	25.78	5.05
p011	24	57,675	450	21	25.33	4.29
p012	24	64,216	465	56	27.17	4.23
p013	24	51,000	509	6	25.96	3.95
p014	24	51,000	279	12	26.58	3.85
p015	24	66,000	369	30	26.21	3.89
p016	27	73,630	1,398	30	26.08	4.28
p017	24	54,000	276	12	26.24	4.37
p018	24	62,942	622	33	25.82	3.89
p019	24	68,172	1,453	71	26.25	4.38
p020	24	55,164	720	9	25.29	3.79
p021	24	69,000	759	57	25.04	3.50
p022	25	72,849	1,194	81	23.37	4.49
p023	24	69,000	852	58	25.26	4.03
p024	24	63,197	513	42	23.81	3.82
p025	24	69,048	1,592	75	24.44	3.67
p026	24	64,887	3,137	37	23.47	3.45
p027	24	72,000	6,524	282	23.02	3.59
p028	24	70,590	264	51	23.41	4.08
p029	26	61,189	1,334	12	26.39	3.61
p030	30	75,000	519	43	28.56	4.04
p031	27	66,836	2,780	28	27.41	3.81
All	765	2,025,631	38,708	1,945	24.81	4.48

Table 3.1: Sample level information for the segmentation dataset. The following information is given for each sample: label (1-4 for individual samples and p001-p031 for pooled samples), number of images (# images) captured, extracted patches from these images (patches ext.), number of sperm cells annotated (# sperm), number of round cells annotated (# round), and the mean length (mean length) and standard deviation of the length (std. length) of the annotated sperm cells. The sperm cells have a mean head length of $24.81\,\text{pixels}$ ($5.67\,\mu\text{m}$) with a standard deviation of $4.48\,\text{pixels}$ ($1.02\,\mu\text{m}$)

Label	Processing	# views	avg. # gt.	avg. # det.	# annotated frames		
					125	256	512
d01		8	-	414.5 (±94.0)	-	-	-
d01	3x dilution	8	41.9 (±11.9)	39.4 (±9.7)	7	-	1
d05		8	-	596.4 (±29.3)	-	-	-
d05	8x dilution	8	-	98.1 (±6.9)	-	-	-
d05	Swim-up	10	21.5 (±4.6)	20.7 (±3.2)	10	-	-
d16	Swim-up	10	-	59.4 (±13.2)	-	-	-
p02		8	101.7 (±23.7)	103.6 (±18.6)	3	-	1
p03		8	-	143.7 (±13.1)	-	-	-
p04		8	-	118.0 (±21.1)	-	-	-
p06		8	-	303.1 (±26.2)	-	-	-
p06	3x dilution	8	-	101.9 (±19.6)	-	-	-
p07		8	-	238.9 (±32.1)	-	-	-
p07	2x dilution	8	-	119.0 (±10.9)	-	-	-
p08		8	-	81.3 (±16.4)	-	-	-
p09		8	98.7 (±0.0)	153.7 (±52.3)	-	-	1
p10		8	-	63.9 (±11.4)	-	-	-
p11		8	-	112.0 (±17.9)	-	-	-
p12		8	-	69.6 (±10.3)	-	-	-
p13		8	-	204.2 (±22.7)	-	-	-
p13	3x dilution	8	-	75.6 (±8.9)	-	-	-
p14		8	-	29.4 (±2.9)	-	-	-
p15		8	47.8 (±4.4)	44.3 (±5.3)	7	1	-
p17		8	-	178.3 (±21.1)	-	-	-
p18		8	-	88.2 (±14.7)	-	-	-
p19		8	-	176.9 (±19.2)	-	-	-
p20		8	-	71.1 (±24.0)	-	-	-
p21		8	25.0 (±4.5)	22.8 (±4.6)	8	-	-
p22		8	122.0 (±0.3)	119.9 (±14.8)	2	-	-
p23		8	-	96.0 (±8.3)	-	-	-
p24		8	-	152.5 (±11.7)	-	-	-

Table 3.2: Tracking dataset information for each sample: label specified as either donor (d) or pooled (p) sample, sample processing (processing), number of views captured (# views), average number of ground truth sperm cells annotated per view (avg. # gt.), average number of automatically detected sperm cells (avg. # det.), and number of views annotated for 125, 256, and 512 frames (# annotated frames). Notice that some samples appear multiple times since capturing was conducted after multiple types of processing. The column "avg. # det." was generated by applying our automatic sperm cell head detector [31] to each frame of each view and counting the resulting detections

Chapter 4

Method

In this chapter we introduce and describe the methods used in the three studies.

4.1 Deep learning

Deep neural networks (DNN) have been studied since the introduction of backpropagation [40] in 1986. However, the methods suffered from being too slow to train on large amounts of data. Since then, hardware development have increased the computational power and storage capabilities yielding training speedup and bigger datasets. The training process gained a speedup factor of up to 100 from using highly optimised software written for and executed on massively parallel hardware such as graphics processing units (GPUs) instead of using traditional sequential code executed on central processing units (CPUs) as described by Ciresan, Meier, and Schmidhuber [3]. This improvement made it possible for DNNs to be trained on bigger datasets within reasonable time. Ciresan, Meier, and Schmidhuber [3] reported training times in the orders of hours or days and relative improvements of the state-of-the-art (at the time of writing) of 30-70% on a wide array of classification datasets.

The ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC2010) [41] dataset was the first large scale image classification dataset to revolutionise deep learning research. It contained 1.2 million images for training, 50,000 images for validating, and 150,000 images for testing. All three subsets had data from 1,000 hand-labelled categories. Krizhevsky, Sutskever, and Hinton [19] were the first to show the power of training a convolutional neural network (CNN) [22] on the imageNet dataset using GPUs. This is by many considered the break-through study inspiring researchers to venture into the field of deep learning. Over the years a wide array of well-known deep CNNs were developed based on the ImageNet dataset such as VGG Net [44], GoogLeNet [48], and ResNet [15].

4.2 The classification task: Logistic regression

In our work we focus on supervised learning for binary classification in which we assume a set of N input data points X, and the corresponding class labels Y from an unknown but fixed distribution $p(X, Y)$. The task is to learn a model $h(\cdot)$ that can be used to predict Y for newly observed X such that $Y = h(X)$. A commonly used assumption is that the data points are independent and identically distributed (i.i.d).

One possible way of solving it is using logistic regression [1], where the likelihood P that $h(\mathbf{x}) = y$ given that $h(\mathbf{x})$ captured the unknown distribution is:

$$P\left(Y|X\right) = \begin{cases} h(\mathbf{x}) & \text{for } y = 1 \\ 1 - h(\mathbf{x}) & \text{for } y = 0 \end{cases} \tag{4.1}$$

Given our assumption of the data points being i.i.d., the following factorisation step holds:

$$P\left(Y|X\right) = \prod_{n=1}^{N} P\left(y_n|\mathbf{x}_n\right) \tag{4.2}$$

Using the logarithm, which is a monotonically increasing function, we optimise the same objective:

$$\log\left(P\left(Y|X\right)\right) = \log\left(\prod_{n=1}^{N} P\left(y_n|\mathbf{x}_n\right)\right) \tag{4.3}$$

$$= \sum_{n=1}^{N} \log\left(P\left(y_n|\mathbf{x}_n\right)\right) \tag{4.4}$$

Maximising the log-likelihood is the same as minimising the negative log-likelihood. Furthermore we wish to obtain an average loss resulting in:

$$-\frac{1}{N}\log\left(P\left(Y|X\right)\right) = -\frac{1}{N}\sum_{n=1}^{N} \log\left(P\left(y_n|\mathbf{x}_n\right)\right) \tag{4.5}$$

Given our binary label space we can substitute $P\left(y_n|\mathbf{x}_n\right)$ with its definition in 4.1 yielding the final formulation of the average negative log-likelihood loss function:

$$-\frac{1}{N}\log\left(P\left(Y|X\right)\right) = -\frac{1}{N}\sum_{n=1}^{N} \left[y_n \log\left(h(\mathbf{x}_n)\right) + \left(1 - y_n\right)\log\left(1 - h\left(\mathbf{x}_n\right)\right)\right] \tag{4.6}$$

We

In this thesis we use a CNN model $h_\theta(\mathbf{x})$ with parameters θ as the hypothesis model. The loss function is minimised by computing the derivatives of $h_\theta(\mathbf{x})$ and changing the parameters accordingly using an optimiser such as gradient descend.

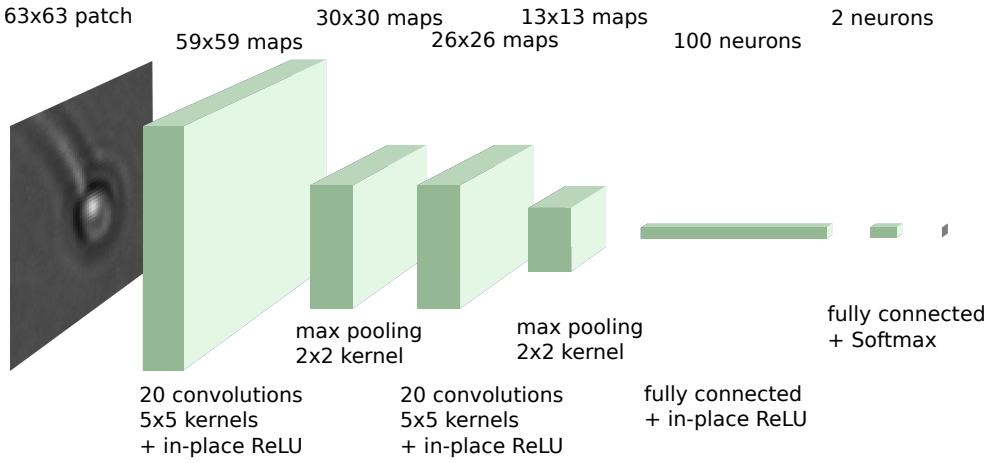


Figure 4.1: Illustration of our 2-conv CNN for performing binary classifying of an image patche into the two classes: background (0) and sperm cell head (1).

4.3 Convolutional neural networks

Neural networks are complex models with a high number of parameters. This combination make the networks difficult to train in practice. Therefore, when designing a neural network we need to constrain the network while maintaining a level of complexity sufficient for representing the unknown distribution in our data.

Given the extent of research within CNNs and the widespread knowledge of the topic I will adhere from describing the methods in detail and instead give a brief introduction to CNNs inspired by Goodfellow, Bengio, and Courville [11] and our work in the first study. A thorough description of deep learning and CNNs in particular can be found in Goodfellow, Bengio, and Courville [11].

In our first study we developed the 2-conv network seen in Figure 4.1. Based on this CNN we explain the components of a traditional CNN model.

A CNN is a neural network with at least one convolution layer [22] applying the convolution operation to its input and optimising the convolution kernel. CNNs are often used as models when there is a grid-like topology such as time series and images as described by Goodfellow, Bengio, and Courville [11]. We exploit these data-dependencies using the convolutional layer to learn patterns in the data while maintaining a minimal number of parameters through weight sharing. Translational equivariance is one of the key properties of the convolution layer: A pattern can be identified irrespective of its location in the input data. The 2-conv network operates on images, where the pixels of the image resembles a grid-like topology. During optimisation, the convolutions learn to recognise features such as the edges around the sperm cell head and the tail irrespective of their location in the image.

In our 2-conv network we have two consecutive “convolution groups” of layers consisting of a convolutional layer, a rectified linear unit (ReLU), and a max pooling layer. The convolution operation is a linear transformation of its input. By using the

ReLU [14] we non-linearly transform the output of the convolution to increase the complexity of the model.

The max-pooling layer [53] summarises its input by performing the max function over small rectangular regions of the image. Other pooling functions can be used such as the average and L_2. The pooling operation gives invariance against slight transformations in the input. In particular, the max pooling layer captures the most prominent feature within a region. Therefore, features learned by the network can be offset slightly yet still recognised by subsequent layers. For example, it is important for the 2-conv network to recognise the round shape of the sperm cell. The shape is formed by combinations of gradients perpendicular to the edge of the sperm cell head. Pooling allows the network to recognise the round shape despite slight distortions of the head.

We compute non-overlapping max pooling operations in the 2-conv network causing a subsampling of the input. However, both the density and region size of the pooling operations can be specified. Subsampling the input decreases the number of operations in subsequent layers making the networks faster in practice. The speed-up is minimal for a small classification network like 2-conv. However, the dimensions of the full images in our segmentation dataset are 1920×1440 pixels, and the addition of subsampling pooling made us able to segment the full image using CNNs in practice.

The 2-conv network has two fully connected layers after the two convolution groups. These layers summarise the existence of features learned by the convolution layers by computing a weighted sum of all elements in the input for each element in the output. Within CNNs these layers are also called the classification layers. The output size should be scaled according to the amount of variation within the classes of the dataset. For example, our 2-conv network needs to recognise sperm cells despite varying appearance due to focus offset and head and tail orientation. Given the linear nature of the fully connected layer, we place a ReLU to increase the complexity of the model.

Finally, we use the softmax layer to compute output probabilities for each of the two classes.

In our first study, we defined the 2-conv-full-up network illustrated in Figure 4.2 conducting dense binary pixel-wise classification of our input images of 1920×1440 pixels. Having described the layers of the 2-conv network we briefly describe the conversions needed to allow dense pixel-wise prediction as described by [24].

Convolution and pooling layers can be applied to images of arbitrary sizes without parameter modification, whereas fully connected layers are input-size dependent. The fully convolutional layer [24] is a convolution layer able to substitute the fully connected layers of the CNN. The two layer types compute identical operations if the number of convolution kernels match the number of fully connected outputs and the size of the convolution kernel matches the size of the input to the fully connected layer. Therefore we can substitute the fully connected layer with a fully convolutional layer (a convolution layer matching the specifications mentioned above). The 2-conv network has a fully connected layer with 100 output elements connected to

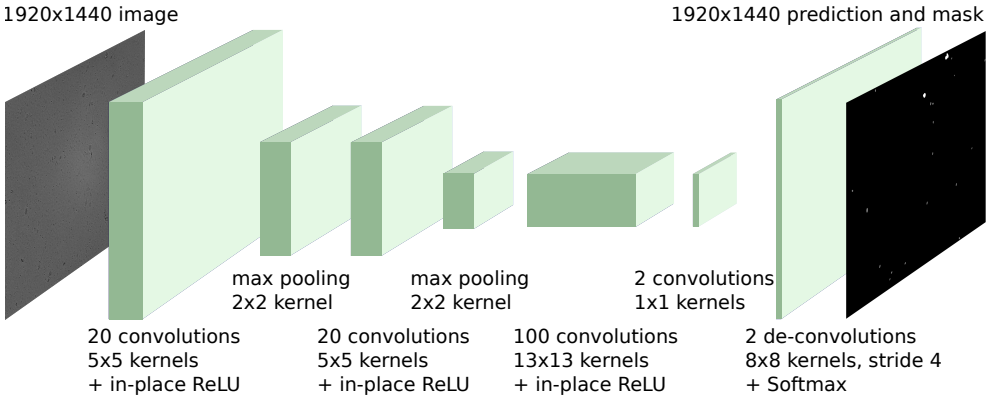


Figure 4.2: Illustration of our 2-conv-full-up CNN for performing dense pixel-wise binary classification on a full size image outputting a full size.

the output of a max pooling layer with size 13×13. We substitute this layer with a convolution layer having 100 kernels of size 13×13 and therefore living up to the specification of the fully convolutional layer. Similarly, the subsequent fully connected layer is substituted by a fully convolutional layer removing the dependency on input image size. The resulting network is called 2-conv-full in our first study.

The 2-conv-full network accepts images of arbitrary size as input and outputs probability maps. However, the output of the network is down-sampled due to the use of max pooling layers. In order to achieve pixel-wise prediction we introduce the transposed convolution layer (also called de-convolution) as described by Long, Shelhamer, and Darrell [24]. The transposed convolution operation is a matrix multiplication conducting the inverse of the convolution. By appropriately initialising the size, stride, and weights of the transposed convolution the layer conducts upscaling of its input resulting in a dense pixel-wise prediction map. The 2-conv-full network is down-sampled by a factor 4. In order to upscale the output by a factor 4, we need to use kernels of size 8×8 and stride 4. The transposed convolution layer is inserted between the last fully connected layer and the softmax layer.

4.4 Manual motility analysis and grading

Manual motility analysis consists of assessing the motility grade of individual sperm cells in order to obtain a motility grade distribution.

Motility grading was previously defined based on general speed of progression [29, 50]: A) Progressive motility ($\geq 25\,\mu\mathrm{m/s}$), B) Slow progressive motility ($5 - -25\,\mu\mathrm{m/s}$), C) non-progressive motility ($< 5\,\mu\mathrm{m/s}$ and tail beat), and D) immotility (no tail beat). The use of general speed of progression made it difficult for the human eye to differentiate between the four motility grades during manual motility analysis as described by Cooper and Yeung [5]. Therefore, grade A and B were merged and the motility grading was verbally defined as according to World Health Organization,

et al. [51, Section 2.5.1, pp. 22]:

- "Progressive motility (PR): spermatozoa moving actively, either linearly or in a large circle, regardless of speed.
- Non-progressive motility (NP): all other patterns of motility with an absence of progression, e.g. swimming in small circles, the flagellar force hardly displacing the head, or when only a flagellar beat can be observed.
- Immotility (IM): no movement."

Thorough guidelines for how to conduct manual motility analysis are described by World Health Organization, et al. [51, Section 2.5, pp. 21-26]: We briefly summarise the steps in the manual motility analysis:

1. Take two aliquots of $10\,\mu\mathrm{L}$ and place the droplets on a glass slide
2. Cover the droplets with a $22 \times 22\,\mathrm{mm}$ cover slip avoiding air bubbles to get a suspension chamber approx. $20\,\mu\mathrm{m}$ deep
3. Place the glass slide on a heated ($37\,^{\circ}\mathrm{C}$) microscope stage
4. Assess whether the two aliquots look similar. Go to step 1 if this is not the case.
5. Allow the sample drift to settle
6. Count/estimate the motility grade of 200 sperm cells in each aliquot
7. Check if the results are within the defined margins of agreement based on the mean and difference of the most frequent motility grade. Go to step 1 if this is not the case. For more details on the margins of agreement see World Health Organization, et al. [51].

4.5 Sperm kinematic metrics

In this thesis we employed a set of kinematic metrics measured for quantifying the motion of a sperm cell. These metrics are globally agreed upon and used by CASA systems and researchers within the field of motility analysis as described by World Health Organization, et al. [51].

Figure 4.3 shows an illustration of the basic kinematic metrics and the variables used for computing the metrics: curvilinear velocity (VCL) is the velocity along the curvilinear path, average path velocity (VAP) is the velocity along the average path, straight-line velocity (VSL) is the velocity from the first to the last track point, amplitude of lateral head displacement (ALH) is the distance between the average and curvilinear paths, and mean angular displacement (MAD) is the average angle between the velocity vectors for two consecutive time steps over time. Based on

the mentioned parameters, a set of descriptive variables can be computed: linearity (LIN, VCL/VSL) is the degree of linear movement along the curvilinear path, wobble (WOB, VAP/VCL) is the degree of side-movement, straightness (STR, VSL/VAP) is the degree of linear movement along the average path, and beat-cross frequency (BCF) is the rate at which the curvilinear and average paths cross.

The motility gradings are defined from general speed of progression, linearity of movement, and tail beating. General speed of progression can be determined by two velocity metrics VSL, and VAP. Linearity of movement is measured by LIN. Tail beating cannot be measured from the sperm cell head position. However, there is no consensus on how to compute VAP in practice whereas VSL is clearly defined. Therefore, we used the VSL, and LIN (VCL/VSL) metrics in studies 2 and 3.

Note that the metrics described are highly dependent of frame rate [7, 2] limiting the application of CASA based findings to CASA systems with similar frame rate.

4.6 Conversion of metrics to motility grade

Sperm kinematic metrics are known to correlate with fertilisation rates [16]. However, we need to convert the kinematic metrics to motility grade when using CASA systems for motility analysis in clinical practice.

The conversion has traditionally been based on gating/thresholding of metrics such as VSL, VCL, and STR and computing the percentage of cells within each gate [29, 5, 12].

Recently, Goodson et al. [12] developed the CASAnova algorithm for classifying motile cells based on VAP, VSL, VCL, ALH, and BCF into five classes with an accuracy of 89.9%. Their model consists of a decision tree with four support vector machines trained from a dataset of tracks with manually classified labels. Their analysis was conducted at a frame rate of $60\,\mathrm{Hz}$ whereas the XC10 captures videos at $25\,\mathrm{Hz}$, therefore we cannot directly use their model due to the metrics' dependency on frame rate.

In our third study we follow the work of Cooper and Yeung [5] and base our motility grading on thresholding of VSL.

4.7 Multi-target tracking

Multi-target tracking is the task of tracking multiple individual objects over time which is required for conducting motility analysis. The task is a concern and issue in CASA systems due to collisions between sperm cells causing incorrect tracking of cells [17].

Recently, Smal and Meijering [46] evaluated a set of nine multi-target trackers for particle tracking in microscopy. The trackers were evaluated on a set of scenarios with varying type of motion, density of objects, and rate of missing and spurious detections.

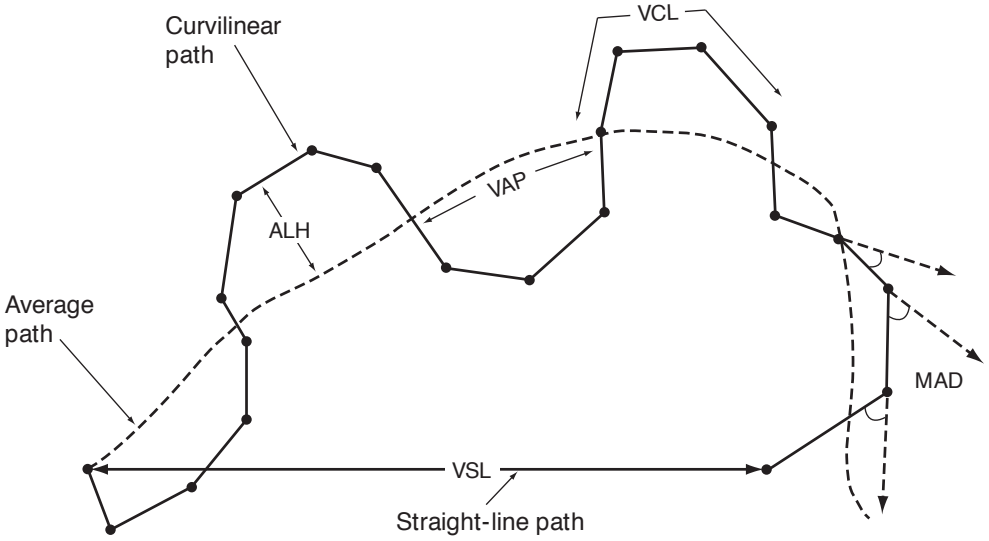


Figure 4.3: Illustration of sperm kinematic metrics. The figure is reprinted with permission from the World Health Organisation [51, Fig. 3.3, pp. 138, Copyright World Health Organisation (2010)].

The Noniterative Greedy Algorithm for Multiframe Point Correspondence (NGA) [46, 43, 18] was consistently amongst the top scoring algorithms of the challenge. Therefore, we used this multi-target tracking algorithm (linker) to link detected sperm cells between consecutive frames.

The linker uses a sliding window approach sequentially linking frames using a frame buffer of size n. At frame i the linker has a set of tracks and a set of detections for frame $i + 1$. The linker considers a set of track extensions between detections from frame $i-n+2$ to detections in frame $i+1$, then $i-n+3$ to $i+1$ until reaching i to $i + 1$. The individual cost of each of possible assignments is computed based on a mixed-motion model defined as a weighted sum of linear and random motion. The assignments are made based on a greedy minimisation of total assignment cost.

In order to use this method, a set of parameters need to be defined. These parameters are: n, diffusion gate, directed motion gate, minimum track length, initial speed of new tracks, cost of missing detection, cost of track death, and cost of birth.

These parameters are highly application dependent. Our choices of parameters are described in study 2 [30].

Chapter 5

Summary of studies and results

5.1 Study I

In our first study [31] "Convolutional Neural Networks for Segmentation and Object Detection of Human Semen" we investigated how to use CNNs for detecting and segmenting sperm cells in bright field microscopy images from the XC10 of human semen samples. The method studied comprised of two steps: pixel-wise segmentation and conversion of said segmentation to objects. CNNs were used for conducting the pixel-wise segmentation, and connected components and filtering based on the component area was used for converting the segmentation to objects. We based the studies on the segmentation dataset described in Section 3.2.

We designed three types of CNNs able to train from one of three possible setups each: an image patch and a ground truth label, a full image and a down sampled label image, or a full image and full label image. The networks trained from patches were regular CNNs whereas the networks trained from full images were FCNs. All networks and an existing baseline sperm detection algorithm described by Ghasemian et al. [10] were trained and tested on the segmentation dataset and compared with each other.

We found that training from full images gave the highest pixel-wise accuracy, precision and recall, and that deeper networks conducting down-sampling as part of the filter bank construction needed up-sampling layers to compensate for the loss of information. Networks with relatively few parameters (20 convolutions per convolution layer) performed almost the same on the train and test dataset whereas a network with increased parameters overfitted to the train dataset though it still performed best on the test dataset.

The baseline method achieved a remarkable difference in performance between the train and test dataset despite the fact that the thresholding area was the only parameter trained. This could indicate a difference in the distribution of variation between the splits of the dataset, which the CNNs had no problem capturing. All our CNNs outperformed the baseline method both on segmentation and object detection performance. Our best performing network achieved a mean intersection-over-union

of 0.7387, 93.87% precision and 91.89% recall on the test dataset. This network had an average prediction time (per image in the dataset) of 0.364 seconds using a single Titan X GPU.

The main contributions of the study were the definition of CNNs and FCNs for pixel-wise segmentation of small objects (sperm cells) in large images and how to train these networks best for both segmentation and object detection.

5.2 Study II

In our second study [30] "Estimation of motility distribution: A study of linearity for human sperm motility analysis" we investigated how to conduct an unbiased estimation of the motility distribution of a population from a subset of observations. The study was based on using the linearity metric for describing the motility distribution of human sperm cells, and we used the tracking dataset described in Section 3.3 for testing the hypotheses of our study. Furthermore, we introduced a fully automatic detection and tracking system based on [31] and the NGA tracking algorithm.

First, we studied metric behaviour, object interaction, and spatio-temporal cell selection to uncover dependencies leading to biases in linearity measurements. Hypotheses for each of the topics were investigated and the following observations were made and 7 guidelines were proposed:

- The linearity metric depends on the observation duration. Therefore, the observation duration should be pre-defined and a potential clinical application should be validated with the specific pre-defined observation duration.
- The variation in linearity measured for a sperm cell depends on the duration of the observation. The minimal variation was observed from a duration of at least 2 seconds.
- Only sperm cells tracked for the full duration should be included in the analysis due to the linearity dependency on duration described above.
- Only cells starting at a location further away from the edge of the view than a distance d should be included. d depended on the duration and was based on the movement speed of the sperm cell population analysed.
- Cells with low motility collided less with other cells than cells with high motility. Therefore, the tracker needs to be capable of handling cell collisions, though the linearity of the single cell was not affected by collisions.
- No global linearity drift was observed during 20 second videos.
- Aggregation of motility statistics across multiple views improved the estimation of the global motility.

Second, we compared linearity distributions computed on the annotated ground truth tracks (theoretical) and on tracks generated using the automatic method introduced (practical). By following the proposed guidelines we achieved nearly identical theoretical and practical linearity distributions.

The seven guidelines proposed based on observations of (real-world) data was the main contribution of the study.

5.3 Study III

In our third study [32] "Evaluation of a new integrated and fully automated system for sperm motility analysis" we evaluated the automatic motility analysis introduced in our second study [30] in combination with the XC10 in a clinical setting. As part of the study we developed the third dataset (Section 3.4) consisting of data from manual and automatic motility analysis of 77 human semen samples, of which 53 samples were analysed with the automatic method more than an hour after the manual analysis was conducted, and 24 samples were analysed with less than an hour between the two analyses.

The evaluation consisted of a comparison between manual and automatic motility analysis, estimation of thresholding parameters for converting the sperm kinematic parameter straight-line speed (VSL) to motility grade (AB, C, and D), and an investigation of the variation in automatic motility read-outs. Bias and variation was estimated using Bland-Altman analysis/plots, and we conducted both 1- and 2-second analysis for all experiments.

First, we conducted a comparison between manual and automatic analysis using the WHO defined AB threshold ($5\,\mu\mathrm{m/s}$) revealing a significant bias between the two analyses.

Second, we optimised the AB and D thresholds to achieve minimal (non-significant) bias. With these thresholds we likewise achieved a very good linear correlation between manual and automatic read-outs of AB% and D%. These thresholds were used throughout the rest of the study.

Third, we investigated the temporal, intra-aliquot, variation in automatic motility read-outs We observed a low temporal variation in read-outs, though the temporal variation was slightly lower in 2- than 1-second analysis. Intra-aliquot variation was slightly higher than temporal variation.

Fourth and last, we investigated the inter-aliquot variation in both manual and automatic motility read-outs. The variation of manual analysis was acceptable (close to $\pm 10\,\%$ as according to [17]) whereas the variation of automatic motility analysis was considerably higher. The variation was caused by a subgroup of samples having higher motility in the first of the two aliquots measured. The variation in automatic analysis was however lower than for manual analysis when only including samples adhering to the agreement between repeated measures as employed in the manual analysis.

In the study we achieved different optimal VSL thresholds for 1- and 2-second analysis indicating a relationship between the two. However, when investigating temporal variation of AB% we observed no significant bias between 1- and 2-second analysis.

The main contributions of the study were the optimised thresholds for converting the kinematic parameter VSL into motility grade resulting in unbiased results compared with manual motility analysis and the investigation of variation from read-outs of the automatic method.

Chapter 6

Discussion and future work

In this chapter we discuss the results presented in our studies and present ideas for future work.

6.1 Detection of sperm cells

From our study on detection of sperm cells [31] we obtained a method capable of segmenting and detecting sperm cells fast and reliably, though both precision and recall achieved by the method could be improved. By combining the detection method with the NGA linker, tracking sperm cells, and computing linearity distributions we obtained linearity distributions almost identical to the theoretically possible distributions. This indicates that our detection method in combination with the linker is sufficient for solving the task of estimating motility distributions of human semen samples despite not achieving perfect precision and recall performance.

The CNNs were developed specifically for segmentation of human sperm cells in our study. However, the CNNs could be used for segmenting other objects of similar size. This would "only" require re-training on a suitable dataset. CM develops systems for counting and quantifying parameters for a wide array of biological cell types. These studies can help both CM and others to develop systems capable of segmenting small objects.

Our choice of segmentation method combined with the simple conversion to objects has one big drawback: Overlapping and adjacent sperm cells are recognised as one big cell. The issue was non-existent when working on the segmentation dataset due to a lack of overlapping and adjacent sperm cells. However, upon studying the tracking dataset we became aware of the severity of the issue when detecting moving and colliding rather than fixated cells. There are multiple ways of possibly fixing the drawback in the method. One approach is to change the post-processing of the segmentation. This would require a study of how to correctly split merged sperm cells based on the probability output of the CNN. The second approach is to include data for adjacent sperm cells in the training dataset, annotate a separation line between adjacent sperm cells, and train the CNNs to recognise the separation line between the

sperm cells correctly. This approach was used by Ronneberger, Fischer, and Brox [39] for training the U-net to recognise separation borders between various (larger) biological cells. The third approach is to design the CNN to solve the instance-aware (semantic) segmentation task. This task has received a lot of attention recently resulting in multiple new instance-aware CNNs [6, 52, 36, 35].

6.2 Motility estimation and tracking of sperm cells

In our second study [30] we thoroughly investigated how to conduct unbiased estimation. One of the key points from the study was that we obtain minimal variation in linearity using 2-second observations. A study by Mack, Wolf, and Tash [25] found that curvilinear velocity (VCL) and VSL means (LIN is the ratio between these metrics) were stable for 5 and 7 trackpoints at 30 frames per second, respectively, contradicting our observation. One key difference between the two studies is, that Mack, Wolf, and Tash [25] based their study on five "representative" sperm cell tracks whereas our studies were based on analysis of several thousand sperm cell tracks.

Currently, the WHO recommends using 1-second analysis [51, Section 3.5.2.2, pp. 138], which reportedly should be sufficient for achieving accurate results. When summarising linearity metrics for all sperm cells in the semen samples, we only observed little difference between the 1- and 2-second analysis in our study [30]. This indicates that the limited variation from 1-second analysis is sufficient for obtaining accurate motility estimates.

According to our reported guidelines we only include sperm cells starting inside a specified field of view. This field of view decreases with increasing observation duration leaving fewer samples for inclusion. Furthermore, we only include sperm cells tracked for the full observation duration. In practice it is harder to track sperm cells for 2 than 1 seconds resulting in fewer tracks living up to the restriction of being traced for the full duration. To summarise, these two requirements cause us to exclude more semen samples when using 2- than 1-second analysis. Therefore we would have to analyse more views using a 2-second analysis than using a 1-second analysis. This observation was very clear in our third study [30] where we had to exclude several semen samples for the 2-second analyses conducted. The choice of whether to use 1- or 2-second analysis depends on the purpose of the analysis. 1-second analysis should be sufficient for routine motility analysis of semen samples. 2-second analysis is preferred when we wish to obtain an accurate estimate of individual sperm cell traits such as when researching effects in motility caused by chemicals [38].

6.3 Kinematic metrics to motility grading

In our third study [32] we compared manual and automatic motility analysis and optimised the VSL thresholding parameters used for converting VSL to motility grading

achieving comparable grading results with no bias. In other words we optimised the thresholds based on a set of semen population statistics measured on two different aliquots of the same samples. By doing so we include possible side effects of sample handling and differences in measurements by the two methods. This optimisation method requires a significant amount of samples given the large variation in motility between semen samples and the practical issues related to sample handling. Currently, we only optimise the thresholds based on 24 and 20 samples for 1- and 2-second analysis, respectively. We will have to collect more data for a higher amount of samples to achieve a confident identification of thresholds.

A different way of optimising the thresholds is to look at the motility grading of the individual sperm cell eliminating the side effects of sample handling. Trained laboratory technicians could manually grade the several thousand sperm cell tracks of our tracking dataset. Thereafter, we could study the relationship between sperm kinematics and motility grade for videos captured by the XC10.

Our automatic motility analysis method relies solely on VSL for motility grading. However, tail beating is one of the differentiating factor between non-progressive (C) and immotile (D) sperm cells during manual motility grading. Therefore it could be beneficial for the automatic analysis to add an estimation of the location of the sperm cell tail. The tail location could likewise be used for indicating general direction of movement when tracking sperm cells. This could improve the tracking especially when solving collision cases. Recall that a minor part of our segmentation dataset already has tails annotated, and thus one could conduct initial studies of tail orientation estimation based on existing data.

6.4 Clinical perspectives

From a clinical point of view we have developed and evaluated an automatic system for conducting motility grading. However, we need to conduct more data collection before being certain that our optimised thresholds are widely applicable and therefore usable in routine analysis as described above. The high intra-aliquot variation observed in automatically obtained read-outs is another concern. During the analysis we identified a subgroup of samples having higher motility in the first than second aliquot most likely caused by drift in the samples. By using the automatic system for conducting the motility analysis we will free up time for the technician. Instead, the technician can focus on sample handling to achieve similar and representative aliquots with minimal drift. However, we need to conduct further studies before concluding whether the intra-aliquot variation poses a problem.

In our studies we have focused on using our detection method for motility analysis. However, detection of sperm cells is also useful for sperm cell concentration and research within SQA. Therefore, the outcome of our studies can also be used to aid researchers in conducting research within SQA.

Chapter 7

Conclusion

In this thesis we presented our studies on automatic motility analysis of human sperm. The studies were based on CM's need for developing image and video analysis for the XC10 and GR's need for automating SQA. The aim of the studies was presented in Chapter 2 and comprised of three questions related to detection and segmentation of sperm cells in less than a second, tracking of sperm cells and obtaining unbiased estimates of heir motility, and evaluating the resulting system for automatic motility analysis of human sperm in a clinical setting. We conducted and presented three studies to answer each of these questions.

First, we presented a method for detecting and segmenting sperm cells using CNNs with high precision and recall sufficiently fast [31].

Second, we recommended a set of guidelines to follow in order to conduct unbiased motility estimation, and we combined our sperm cell detector with an existing tracking algorithm resulting in reliable estimates of motility distributions [30].

Third and last, we evaluated the proposed method for automatic motility estimation by comparing it with manual motility analysis resulting in comparable motility estimates. The system can help researchers within human SQA, thought more data needs to be collected before concluding whether the system can be used in routine motility analysis of human semen samples.

Part II

Papers

Chapter 8

Paper I: Convolutional Neural Networks for Segmentation and Object Detection of Human Semen

Convolutional Neural Networks for Segmentation and Object Detection of Human Semen

Malte S. Nissen[1,2,3,4( )], Oswin Krause[1], Kristian Almstrup[2,3], Søren Kjærulff[4], Torben T. Nielsen[4], and Mads Nielsen[1]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
nissen@di.ku.dk
[2] Department of Growth and Reproduction, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark
[3] International Center for Research and Research Training in Endocrine Disruption of Male Reproduction and Child Health (EDMaRC), Rigshospitalet, University of Copenhagen, Copenhagen, Denmark
[4] ChemoMetec A/S, Allerød, Denmark

Abstract. We compare a set of convolutional neural network (CNN) architectures for the task of segmenting and detecting human sperm cells in an image taken from a semen sample. In contrast to previous work, samples are not stained or washed to allow for full sperm quality analysis, making analysis harder due to clutter. Our results indicate that training on full images is superior to training on patches when class-skew is properly handled. Full image training including up-sampling during training proves to be beneficial in deep CNNs for pixel wise accuracy and detection performance. Predicted sperm cells are found by using connected components on the CNN predictions. We investigate optimization of a threshold parameter on the size of detected components. Our best network achieves 93.87% precision and 91.89% recall on our test dataset after thresholding outperforming a classical image analysis approach.

Keywords: Deep learning · Segmentation · Convolutional neural networks · Human sperm · Fertility examination

1 Introduction

Sperm Quality Analysis (SQA) involves measuring concentration, morphology, and motility [13] of sperm cells. For the application to animal sperm cells, there exist a number of commercial Computer-Aided Sperm Analysis (CASA) systems, such as the Hamilton-Thorne *IVOS-II* and *CEROS-II*[1] and the *Sperm Class Analyzer*[2].

[1] http://www.hamiltonthorne.com/.
[2] http://www.micropticsl.com/products/sperm-class-analyzer-casa-system/.

P. Sharma and F.M. Bianchi (Eds.): SCIA 2017, Part I, LNCS 10269, pp. 397–406, 2017.
DOI: 10.1007/978-3-319-59126-1_33

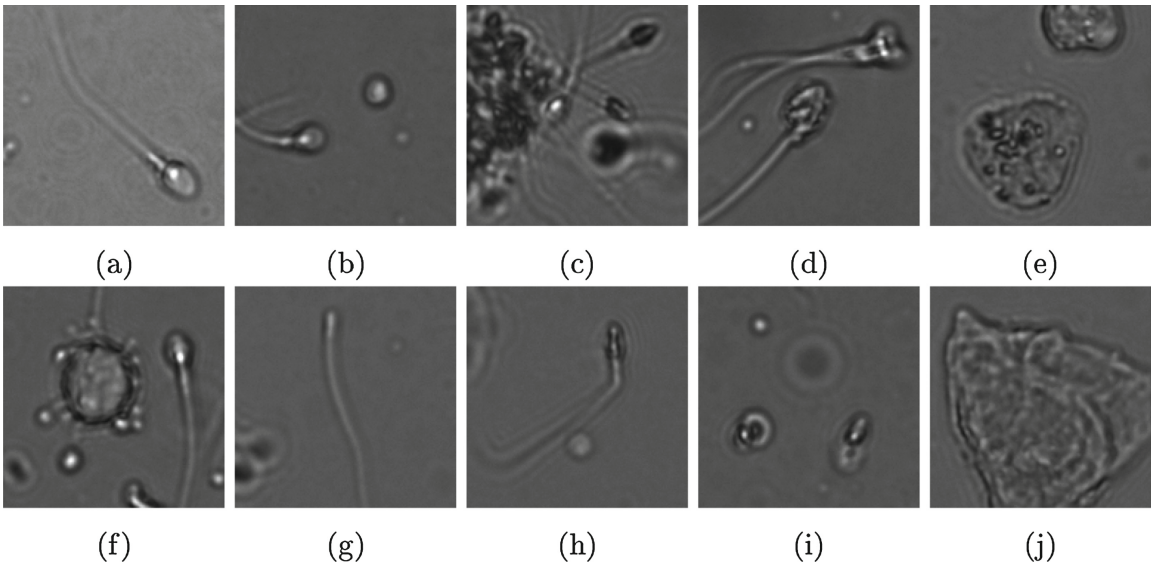


Fig. 1. Examples of debris, variations, and morphological abnormalities: normal sperm cell (a, b), aggregated cells out of focus (c), agglutinated cells (d), round cells (e, f), headless sperm (g), sperm head seen from the side or morphologically abnormal (h, i), circular tails (i), and other types of artifacts and debris (b, f, j).

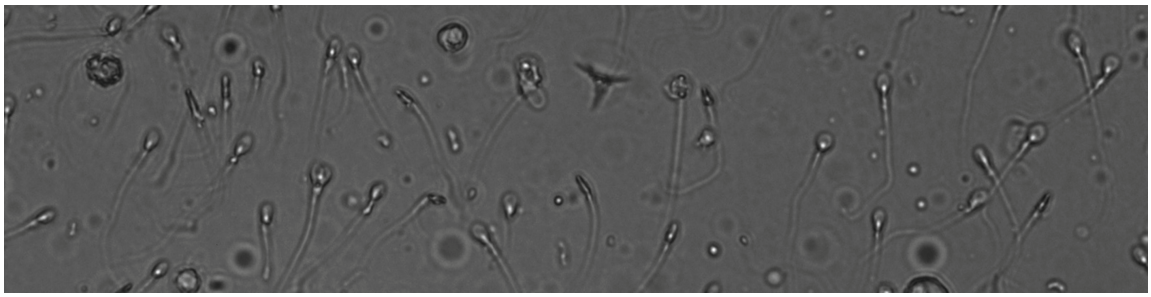


Fig. 2. 1200 × 300 pixel cut-out of image from the dataset

Human semen samples have a significantly lower quality of sperm cells compared to most animals [7], which increases the accuracy demand on the analysis. Moreover, human semen is often cluttered with debris and cells other than normal mature sperms. Figure 1 shows examples of typical debris, variations, and morphological abnormalities of human sperm samples. Figure 2 shows a section of a typical image.

In practice, staining and smearing are often used for preparation of samples to highlight specific properties of the cells [1–4,10], but the sample needs to be in its natural form for motility estimation. This article focuses on the first step of SQA, image segmentation and detection of non-stained human sperm cells as analyzed by Ghasemian et al. [4] and Hidayahtullah and Zuhdi [6]. These algorithms apply classical image analysis techniques to solve the problem. To our knowledge no deep learning techniques have been applied yet.

Our approach focuses on deep convolutional neural networks (CNN) to segment the sperm cells in the image. There are three main challenges in this

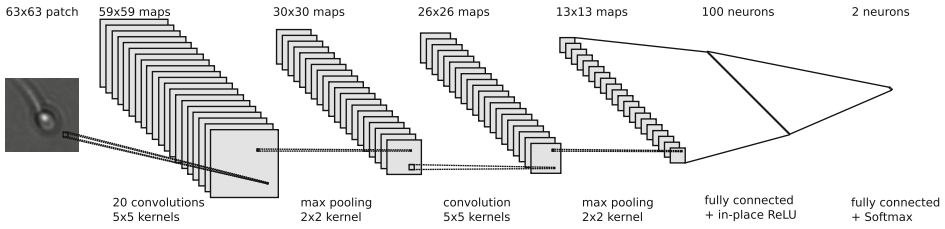


Fig. 3. Illustration of the 2-conv CNN

approach: Firstly, every pooling layer in a CNN reduces resolution by at least 50%; after three layers of pooling, every pixel of the result encodes the information of an 8×8 area of the original image. Secondly, CNNs are often trained on image patches, however there is a huge class imbalance between background and sperm pixels, where sperm pixels are significantly harder to detect. Lastly, we need to cluster the segmentations to objects. Imperfect predictions of the networks often lead to spurious detections, which need to be removed. One way to do this is to use thresholding on the size of clusters, leading to an arbitrary threshold parameter. This parameter needs to be chosen carefully.

We investigate possible solutions to these challenges. While using max-pooling layers is possible without reducing resolution [5], an exponential amount of time in the number of pooling layers is required. This makes it infeasible in practice as the results have to be computed quickly enough to allow video analysis. We follow Long et al. [9] and investigate up-sampling on the output of the CNN during training and testing. Ronneberger et al. [11] proposed a more complex architecture, which we disregard since predictions would be too slow for our application. Further, we compare training on image patches with training on the full images, where class-labels are re-weighted to correct the class-skew.

For comparison we implemented the sperm head detection method proposed by Ghasemian et al. [4]. This method has a similar threshold parameter as our method which has to be adapted for a fair comparison. For this, we propose a way to adapt the thresholding parameters using the product of precision and recall on the final detections.

The paper is organized as follows: Sect. 2 describes the dataset and the CNN architectures used. Experiments are described in Sect. 3. Results are given in Sect. 4 and discussed in Sect. 5. Finally, we conclude in Sect. 6.

2 Method

Dataset. We have constructed a dataset of 765 grayscale images of 35 independent sperm samples. The 35 samples were individually diluted using a solution of Bicarbonate-Formalin (as devised by WHO [13]) to get an appropriate amount of cells in each image (between 2 and 290 sperm cells) and to fixate them. Fixation facilitates sedimentation of the cells to the bottom of the counting chamber, ensuring that all cells are roughly in the same focal plane. In order to have cells both in and out of focus, reflecting the optical variation, Z-stacks of images were

acquired. The images were acquired using an image cytometer with $20\times$ optical magnification and a resolution of 1920×1440 pixels (0.2 µm/pixel). The image intensities have been quantized from 14- to 8-bit images. In each image the intensities where normalized to lie between zero and one.

The images were annotated by experts and registered into two classes: background and sperm cells. Round cells form an important part of the background and were therefore also annotated. The tip of the head and the neck point was registered for each sperm cell while the circumference was annotated for each round cell. Pixel-segmentation ground truths are generated by creating an ellipse at the center of each sperm cell head with radius $r_1 = \frac{1}{4}l_{cell}$ and $r_2 = \frac{2}{3}r_1$ where l_{cell} is the length of the cell head.

We split the samples into 70% train and 30% test data based on stratified sampling on the average number of sperm cells in the full images of each sample. This ensures that images from the same sample are part of the same split as they contain correlated data. Hence, one sample being part of testing data is never represented in the training data.

From the training dataset we generated an additional dataset of extracted patches from the images using the annotated classes. This patch dataset contains 63×63 pixel patches which are labelled by their ground truth in the center pixel. The size of the patches is chosen to allow the entire head, which is typically 25 pixels long, and a small part of the tail to be included. From each image, we extract up to 3,000 patches, split into 40% sperm cells, 40% background and 20% round cells. The numbers were chosen to cover the variety of debris in the background class (round cells contribute a lot to the variability of the background). Random rotation and flipping is applied before extracting each patch. Table 1 shows statistics for the resulting datasets. Note that the dataset contains a total of 38,708 sperm cells of which 23,997 are included in the train set and 14,711 are included in the test set.

Table 1. Data statistics

Statistic	Train	Test	Total
Images	540	225	765
Sperm cells	23,997	14,711	38,708
Patches	1,424,341	601,290	2,025,631

Networks. We define seven networks to test against each other. The first network is called *2-conv*. It is defined for input patches and illustrated in Fig. 3. It is a standard CNN with two convolutional, ReLU and max-pooling layers followed by two fully connected layers separated by another ReLu layer and including 50% dropout during training. The network *3-conv* is obtained by adding an additional set of convolution, ReLU, and max-pooling layers. The networks are defined with receptive fields of size 63×63 using 20 filters in each of their convolution layers and 100 filters in their fully convolutional layer.

Table 2. Experiment results m_{IU}, threshold, and m_{pred} for all eight methods

Method	m_{IU}	Threshold	m_{pred} (s)
2-conv	0.6658	200	0.145
2-conv-full	0.7080	200	0.143
2-conv-full-up	0.6805	250	0.143
3-conv	0.6556	200	0.119
3-conv-full	0.6497	150	0.119
3-conv-full-up	0.6661	300	**0.116**
3-conv-full-up-inc	**0.7387**	150	0.364
Baseline [4]	0.5679	400	-

For prediction on the full images, the fully connected layers are substituted with fully convolutional layers as described by Long et al. [9] to allow for faster computation. As each max-pooling layer divides the spatial resolution of the output by a factor of 2 in each dimension, we further perform bilinear upscaling of the network output probabilities to obtain a pixel-wise segmentation.

To compare whether training on full images is beneficial compared to patch-based training, we define the architectures *2-conv-full* and *3-conv-full*, which have the same structure as *2-conv* and *3-conv* in the prediction phase and are trained on full images with the final up-sampling removed. Finally, the architectures *2-conv-full-up* and *3-conv-full-up* also incorporate the bilinear up-sampling into the training process. The networks trained on full images use a receptive field of size 64×64 and the same number of filters[3]. We further add a network *3-conv-full-up-inc* with the same receptive field size but with 64, 128, and 256 filters in the convolution layers and 1024 filters in the fully convolutional layer. We omit the network *2-conv-full-up-inc* due to limitations in the framework used.

When testing the networks, we perform post-processing of the full size output probabilities in two steps: Firstly, we choose the most probable class as output for each pixel. Secondly, we cluster pixel-wise segmentation to objects by computing the 8-neighbourhood connected components and removing components smaller than a threshold t. The value of this threshold is found in Sect. 4.

3 Experiments

The 2-conv and 3-conv architectures have been trained on the patch dataset and tested on the full image dataset, whereas all other networks have been trained and tested on the full image dataset. The outputs of 2-conv-full and 3-conv-full are smaller than the label masks of the full images. We therefore downsample the label masks by factors 4 and 8 respectively. This is done by taking every 4th or 8th pixel corresponding to the center of the receptive field of the output.

[3] The difference comes from the fact that it is easier to define a center-pixel in 63×63 receptive fields.

All networks are trained by optimizing the cross-entropy between the predicted and ground truth label. To compensate for the class skew in the full images during training we re-weight the classes according to their distribution. The weight w_i of class i is defined as $w_i = \frac{1}{n_i \sum_j \frac{1}{n_j}}$ where n_i is the number of pixels belonging to class i. Omitting the re-weighting led to far inferior results classifying everything as background.

The architectures have been trained for 200 epochs using the Adam solver [12] with mini-batches of 256 patches or 1 full image ($1920 \cdot 1440$ "samples"). For training we chose learning rate $\alpha = 0.001$, moment 1 $\beta_1 = 0.9$, moment 2 $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We implemented the networks using Caffe [8], and the experiments have been carried out using a single Titan X GPU.

The baseline method [4] consists of three major steps: Noise reduction, object region detection, and sperm head localization. The method assumes that all sufficiently large object regions are sperm cells and therefore filters out all object regions smaller than a chosen threshold. This threshold is crucial for the performance of the algorithm and needs to be chosen carefully.

On an object level we are interested in finding each sperm cell. For this purpose we use the two measures precision $= \frac{TP}{TP+FP}$ and recall $= \frac{TP}{TP+FN}$, where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. A predicted sperm cell is categorized as TP if it covers more than half the area of a ground truth sperm cell. Each predicted cell can only count as one positive, i.e. a predicted cell covering more than half the area of two sperm cells counts as one true positive and one false negative. We evaluate precision and recall for multiple thresholds on the training data to get a precision-recall (PR) curve for every method. We choose the threshold value that maximizes the product between precision and recall.

Mean intersection over union (mean IU) m_{IU} is used to quantify the pixel-wise segmentation performance as described by Long et al. [9]:

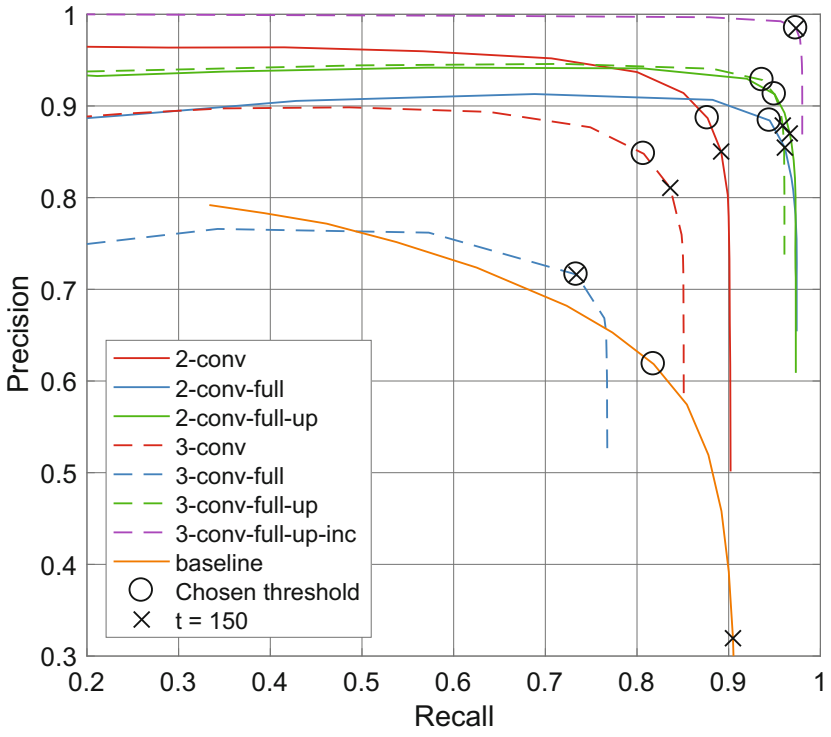
$$m_{IU} = \frac{1}{2} \sum_i \left(\frac{p_{ii}}{\sum_j (p_{ij} + p_{ji}) - p_{ii}} \right)$$

where p_{ij} is the number of occurences of class i predicted as class j. We have chosen this measure since it is invariant to the aforementioned class skew.

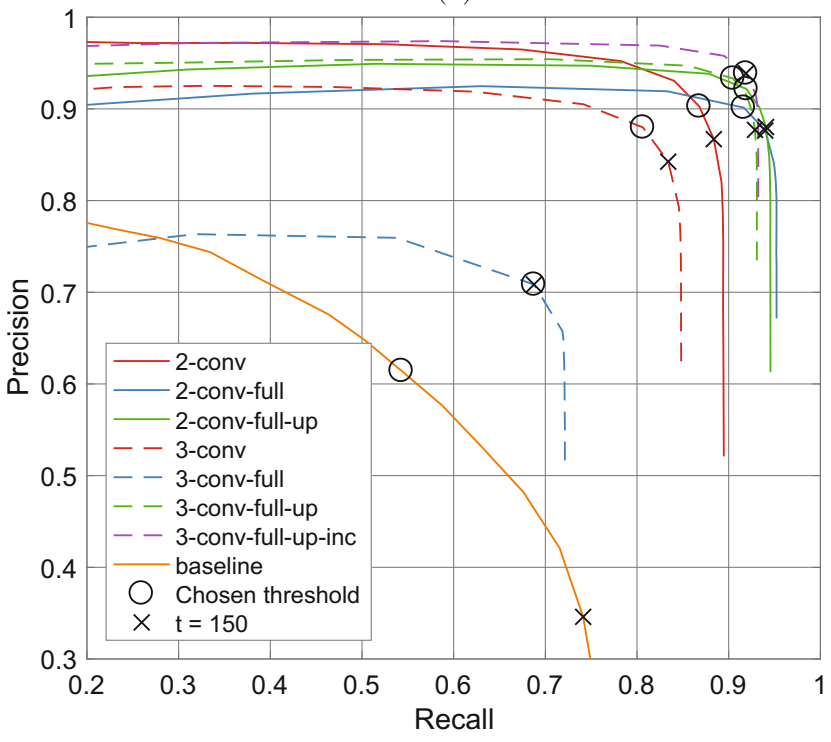
Finally, fast computations is one of the requirements for automatic SQA. We therefore record the execution time of computing a prediction and object removal on all 765 full images and compute the mean execution time m_{pred} per image. Our baseline method implementation is not as optimized as our networks and therefore we omit the results.

4 Results

Results of the mean IU m_{IU}, thresholds found by maximizing the product between precision and recall on training data, and mean execution time m_{pred} for each method are given in Table 2.



(a)



(b)

Fig. 4. Precision-recall graphs for (a) train and (b) test for all networks. 2-conv networks are plotted using fully-drawn lines and 3-conv networks using dashed lines. Same colours of lines indicate same parameters. The black circles indicate the point on each graph that corresponds to the threshold t reported in Table 2, while the black crosses indicate the points for $t = 150$.

Generally when considering m_{IU}, the networks trained on full images with up-sampling perform better than the networks trained on patches. All networks perform better than the baseline method. Training on full images without up-sampling leads to better results for 2-conv-full but worse for 3-conv-full. The 2-conv networks perform better than their 3-conv equivalents. The network *3-conv-full-up-inc* performs best, but it also has a considerably higher execution time $m_{pred} = 0.364$ than the other methods spanning the range of 0.116–0.145 seconds per image.

The results for the object detection are given in Fig. 4. The figure shows the precision-recall graphs for (a) train and (b) test for all methods. The graphs have ends due to the smallest and largest thresholds considered (0–1, 000). We plot 2-conv networks using fully-drawn lines and 3-conv networks using dashed lines. Same colours of lines indicate same parameters. The black circles indicate the point on each graph that corresponds to the threshold reported in Table 2, while the black crosses indicate the points for a threshold of 150. The baseline performs considerably different on the train and test set even though there is no training involved apart from the choice of threshold. It performs considerably worse than our networks except 3-conv-full. The best method is 3-conv-full-up-inc having 93.87% precision and 91.89% recall on the test set using threshold 150. While some overfitting can be seen between training and test, it still outperforms the other methods.

5 Discussion

Our results show that using neural networks is beneficial compared to the classical approach. The large difference in baseline performance indicate that there is a large variation between samples. We believe that we have captured the variation of a sperm cell in our train and test sets, but we have not captured all possible combinations of cells in an entire image. Given our limited number of individual samples, there are some cell concentration differences. The baseline performance difference is likely caused by these cell concentration differences. Our networks are not affected by these differences except to the degree expected from overfitting. All networks except 3-conv-full-up-inc perform almost the same on train and test data whereas 3-conv-full-up-inc is showing clear signs of overfitting. This indicates that our networks are sufficiently complex to cover the variation of the data and that even larger networks are unlikely to generalize better. As we have not used the test set for model selection, we can expect the performance on the test set to be close to the true performance.

Up-sampling has different effects on mean IU and object detection. For mean IU detecting object boundaries is important. As up-sampling is equivalent to blurring it is not beneficial for mean IU when the model is already able to accurately describe the shape of the objects. This can be seen in the difference in its effect on networks with two and three max-pooling layers. We hypothesize that training using up-sampling gives us true predictions with cluster areas closer to the true size of sperm cells. This makes it easier to distinguish sperm cells

from a specific type of debris (Fig. 1b and i) easily mistaken for the head of a sperm cell but having a slightly smaller area.

When omitting up-sampling, there is no general tendency when comparing patch-based and full-image training. For 2-conv networks, full-image training seems to profit from the increased variation in the data while patch-based training profits from the weighting of round cells in the background. This can be seen by the differences in precision and recall for the two methods in Fig. 4.

When we compare the PR-curves, we see that the choice of a fixed threshold can be misleading. It turns out that the ranking of the networks can change depending on the choice of it. However, the chosen thresholds on the training set lead to consistent rankings on the test set in our case. Introducing the threshold and optimizing it leads to far superior results for all networks compared to choosing an arbitrary value. The obtained precision and recall seems reasonable for the purpose of identifying sperm cells in a semen sample, however it needs clinical testing for verification of its performance in practice.

6 Conclusion

In this paper, we have used deep convolutional neural networks for the task of sperm cell segmentation and object detection. In this task, we are constrained by the computation time as well as the accuracy demands, which make it harder to train networks with many pooling layers. To mitigate both problems we explored the use of full image training and up-sampling of the network outputs in order to increase performance. We specifically investigated thresholding on the size of detected components. Choosing the product of precision and recall leads to a robust estimate of threshold parameter. For deeper networks, up-sampling appears necessary to achieve good segmentation and object detection performance. The same does not necessarily hold for more shallow networks.

Our method outperformed a classical image analysis method which can be considered state-of-the-art. Overall the system sensitivity and precision are sufficiently high to be valuable for human sperm analysis systems.

Acknowledgements. This work is partly funded by the Innovation Fund Denmark (IFD) under File No. 4135-00169B. We would like to thank Department of Growth and Reproduction, Rigshospitalet, Denmark, for helping with annotation of our data.

References

1. Bijar, A., Pe, A., Mikaeili, M., et al.: Fully automatic identification and discrimination of sperm's parts in microscopic images of stained human semen smear. Scientific Research Publishing (2012)
2. Carrillo, H., Villarreal, J., Sotaquira, M., Goelkel, M.A., Gutierrez, R.: A computer aided tool for the assessment of human sperm morphology. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, pp. 1152–1157 (2007)

3. Chang, V., Saavedra, J.M., Castañeda, V., Sarabia, L., Hitschfeld, N., Härtel, S.: Gold-standard and improved framework for sperm head segmentation. Comput. Methods Progr. Biomed. **117**(2), 225–237 (2014)
4. Ghasemian, F., Mirroshandel, S.A., Monji-Azad, S., Azarnia, M., Zahiri, Z.: An efficient method for automatic morphological abnormality detection from human sperm images. Comput. Methods Progr. Biomed. **122**(3), 409–420 (2015)
5. Giusti, A., Cireşan, D.C., Masci, J., Gambardella, L.M., Schmidhuber, J.: Fast image scanning with deep max-pooling convolutional neural networks. arXiv preprint (2013). arXiv:1302.1700
6. Hidayatullah, P., Zuhdi, M.: Automatic sperms counting using adaptive local threshold and ellipse detection. In: International Conference on Information Technology Systems and Innovation (ICITSI), pp. 56–61 (2014)
7. van der Horst, G., Mortimer, S.T., Mortimer, D.: The future of computer-aided sperm analysis. Asian J. Androl. **17**, 4 (2015)
8. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint (2014). arXiv:1408.5093
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
10. Medina-Rodríguez, R., Guzmán-Masías, L., Alatrista-Salas, H., Beltrán-Castañón, C.: Sperm cells segmentation in micrographic images through lambertian reflectance model. In: Azzopardi, G., Petkov, N. (eds.) CAIP 2015. LNCS, vol. 9257, pp. 664–674. Springer, Cham (2015). doi:10.1007/978-3-319-23117-4_57
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_28
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint (2014). arXiv:1412.6980
13. World Health Organization and others: WHO laboratory manual for the examination and processing of human semen. World Health Organization, Geneva (2010)

Chapter 9

Paper II: Estimation of motility distribution: A study of linearity for human sperm motility analysis

Transactions on Medical Imaging manuscript
(to be submitted)

Estimation of motility distribution: A study of linearity for human sperm motility analysis

Malte Stær Nissen[1,2,3] ·
Kristian Almstrup[2,3] ·
Oswin Krause[1] ·
Søren Kjærulff[4] ·
Erik Meijering[5] ·
Mads Nielsen[1]

Received: date / Accepted: date

Abstract We investigate how to conduct an unbiased motility estimation based on using the linearity metric for human sperm motility analysis. The three major topics of our investigation are metric behaviour, object interaction, and spatio-temporal object selection. For validation of the problems and solutions we have constructed a dataset of 2,730 semi-automatically annotated (ground truth) sperm cell tracks from 41 videos of 7 independent human sperm samples. The investigation showed that the observation duration should be pre-defined and the motility estimation should be clinically validated for the specific duration. We identified no global drift in linearity in videos analysed for up to 20 seconds. Using observations of 2 seconds maximises the consistency of measurements for each sperm cell. We should only include sperm cells tracked for the full duration to avoid bias towards higher motility. Furthermore, we should only include cells that start within a field of view defined from a minimum distance to the edge of the view based on the observation duration. Sperm cell collisions need to be handled to avoid introducing an implicit selection bias. Finally, aggregating motility statistics across multiple views of the sample improves our estimate of the global motility.

By following these recommendations we were able to achieve very similar motility distributions when comparing theoretical (based on ground truth tracks) and practical (based on automatically detected and tracked sperm cells) estimates for 1- and 2-second analysis. Therefore, we can expect practical motility estimates to reflect the theoretically possible motility estimates on similar data.

Keywords Motility distribution estimation · Selection bias · Observation statistics · Track linearity

This work is partly funded by the Innovation Fund Denmark (IFD) under File No. 4135-00169B, and by ChemoMetec A/S, Gydevang 43, DK-3450 Allerod, Denmark

M. Nissen
E-mail: nissen@di.ku.dk
[1]Dept. of Computer Science, University of Copenhagen, Copenhagen, Denmark
[2]Dept. of Growth and Reproduction, Rigshospitalet, University of Copenhagen, Denmark
[3]International Center for Research and Research Training in Endocrine Disruption of Male Reproduction and Child Health (EDMaRC), Rigshospitalet, University of Copenhagen, Denmark
[4] ChemoMetec A/S, Allerød, Denmark
[5] Biomedical Imaging Group Rotterdam, Erasmus University Medical Center, Rotterdam, the Netherlands

1 Introduction

Multiple machine vision applications use estimation of motion and motility of objects, such as: movement behaviour of zebrafish, behaviour analysis based on hand and finger movements, and motility analysis of human sperm cells [10, 7, 18].The central question in all these applications is how the movement of objects can be traced and described best so as to facilitate subsequent analyses.

In practice we can only observe a spatially and temporally limited view of the population. Based on this limited observation we can compute the desired motility metrics and estimate the motility distribution.

During estimation we need to consider various effects possibly biasing the estimate. Fig. 1 illustrates the spatio-temporal space with the x- and y-axes representing the spatial dimensions and the t-axis representing the temporal dimension. The cube illustrates a sample observation in the spatio-temporal space. The four tracks inside the cube illustrate four basic classes of tracks we encounter: A) objects *fully tracked* within our observation, B) objects present at the beginning of our observation but *exiting* the view before the observation ends, C) objects *passing* through the view thus entering after the observation starts and exiting before the observation ends, and D) objects *entering* and staying within the view until the observation ends. Classes B, C, and D only contain moving objects whereas class A also includes stationary objects. Only moving objects exit and enter the view during the observation. Therefore classes B, C, and D potentially skew the motility distribution inside the view. In some applications objects interact with each other possibly changing their movement and thus also affecting the metrics measured. The

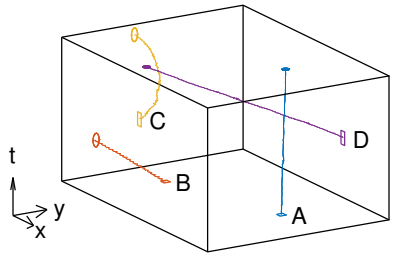


Fig. 1 The tracking problem illustrated by the spatial axes x and y and the temporal axis t. The cube represents a view of objects from a population. The lines inside the cube represent object track groups: fully tracked object (A, blue), exiting object (B, purple), passing object (C, red), entering object (D, yellow). The objects enter and exit the cube at their square and circular markers respectively.

three main topics we investigate are: *Metric behaviour*, how does the metric depend on the spatio-temporal domain? *Object interaction*, how do objects interact? And *Spatio-temporal object selection*, which objects do we include in the analysis to avoid selection bias influencing resulting?

Our application is human sperm motility analysis, which is an essential part of human sperm quality analysis. The analysis is mainly conducted manually as described by the WHO [20], and is consequently very labour intensive and prone to high variation, though the latter can be reduced with the aid of repeated training [6,8]. Here we have automated the analysis with the aim to overcome these problems. Our analysis of motility hence consists of the sperm cells in a semen sample, and the observation is a video of a limited view and temporal duration of the semen sample observed through an image cytometer.

1.1 Human sperm motility analysis

The WHO [20] describes the basic sperm quality analyses. It contains a set of recommendations for using computer-aided sperm analysis (CASA) systems for motility estimation. In short, a small portion of the semen sample is put under a microscope, videos of multiple views are captured, and cells are identified and tracked throughout the videos. Manual assessment of the motility of each sperm cell is done according to the definition of motility in the WHO manual [20, section 2.5.1]:

- "Progressive motility (PR): spermatozoa moving actively, either linearly or in a large circle, regardless of speed.
- Non-progressive motility (NP): all other patterns of motility with an absence of progression, e.g. swimming in small circles, the flagellar force hardly displacing the head, or when only a flagellar beat can be observed.
- Immotility (IM): no movement."

This verbal definition of sperm motility leaves room for interpretation of how to estimate the motility from sperm tracks when going from manual to automatic assessment. The WHO manual briefly describes some of the standard sperm track kinematic characteristics measured by existing CASA systems: curvilinear velocity VCL, straight-line velocity VSL, average path velocity, amplitude of lateral head displacement, linearity LIN, wobble, straightness, beat-cross frequency, and mean angular displacement. Earlier versions of the manual based the motility grading on movement speed rather than linearity. Therefore pre-2010 studies and CASA systems typically base their motility estimates on the cell speed measures [18,3]. We here aim at measuring the percentage of progressively moving cells, and thus we investigate linearity for being able to distinguish between PR and NP/IM cells.

When automating sperm motility analysis several questions are raised which we try to address in the present study. These include:

Sperm cells are easily affected by their living conditions. They gradually experience a decrease in motility after ejaculation, and temperature changes affect their movement. Therefore the motility needs to be estimated within 60 minutes of ejaculation and at a well specified temperature, typically 37 °C, in order to get the most accurate motility estimate. Do we experience a global drift in motility due to these factors?

According to the WHO manual one needs to follow a sperm cell for at least one second for accurately estimating its motility. This is due to the occasionally irregular behaviour of sperm cells (noise). Does linearity depend on the duration for which it is measured, and how long a duration do we need to get consistent motility estimates of a cell?

Sperm cells moving into the view skew the statistics, and therefore motility is in general easily overestimated. How much do entering cells skew the statistics and how do we select cells to get an unbiased motility estimate?

Finally, cell collisions occur even at low sperm cell concentrations. In practice, collisions can be hard to solve correctly. This could cause the tracks of colliding cells to die affecting the motility distribution measured. Do collisions affect the linearity measured on cell level, and is the linearity distribution of colliding cells different from the linearity distribution of non-colliding cells?

1.2 Topics and hypotheses

We split our investigation into the three main topics in order to assess the effect of the three main pitfalls mentioned above: metric behaviour, object interaction, and spatio-temporal object selection. For each of these topics we define a set of hypotheses:

- Metric behaviour:
  1. Global drift: The global drift in motility due to cell exhaustion and temperature changes cannot be observed in videos of 20 seconds
  2. Linearity decrease: Measured linearity decreases with increased observation duration
  3. Metric consistency: The consistency/stability of measured linearity depends on the observation duration
- Object interaction:
  1. Collisions affect cell motility
- Spatio-temporal object selection:
  1. Cell origin: Entering and passing (introduced) cells are biased towards progressively motile compared to fully tracked (origin) cells
  2. Field of view: Requirements on the observation duration cause a selection bias towards slow and immotile sperm cells

Based on the results of the investigation we propose a protocol for analysing a video of human semen. We compare the outcome of applying the protocol to ground truth tracks and tracks obtained from automatic detection and linking of sperm cells in videos for which we have ground truth tracks.

1.3 Structure of the article

The structure of the remaining of the article is: In section 2 we describe the dataset and methods used for analysis. Section 3 describes the experiments and findings related to each hypothesis. We discuss the results and propose an analysis protocol in section 4, and finally we conclude upon our work in section 5.

2 Methods

2.1 Dataset

Our dataset consists of videos of 7 fully anonymised independent samples. Table 1 shows an overview of information about the samples in the dataset. The samples include two donor samples (d01 and d05), and five pooled samples each consisting of semen from three patients. The raw sample was allowed to liquefy and kept at 37 °C until being loaded in a 20 µm glass chamber, after which we captured 8-10 independent views using bright field lighting and 20x optical magnification in the ChemoMetec Xcyto 10 image cytometer[1]. The sperm cell concentration of each sample was estimated using the automatic image cytometry method of Egeberg et al. [5]. The sample "d01" was diluted to a 3x solution using phosphate-buffered saline (PBS). in order to limit the number of cells per view, and the sample "d05" went through a swim-up procedure similar to [14] to purify the sample and leave an overrepresentation of progressively motile sperm cells. Collectively, the samples represent a broad spectrum of samples encountered in a fertility clinic.

Each video consists of 512 greyscale frames of 1920×1440 pixels ($435.6\,\mu m \times 326.7\,\mu m$) captured at 25 frames per second (fps).

Annotation of the full dataset proved to be too time consuming, and thus we only annotated 41 of the views for 125-512 frames. Table 1 shows detailed information about the number of annotated views from each sample, the duration of the annotations, and the average number of cells per view. Notice that we annotated three views for the full duration of 512 frames, one view for 256 frames, and 37 views for 125 frames. The choice of annotation duration was adjusted throughout the annotation process based on time consumption. The annotations were conducted semi-automatically. An initial detection of sperm cells in each frame was conducted using the method of Nissen et al. [13], non-collision nearest neighbour points were automatically connected, collision points were manually solved, missing points and tracks were manually added, and finally all tracks were individually manually inspected for errors and corrected. This protocol took between 2-37 hours to conduct per view depending on the concentration of cells, their movement, and the number of annotated frames.

Table 2 shows a brief summary ground truth track information grouped by the number of annotated frames. Fig. 2 shows a histogram of the ground truth track lengths. Notice that the y-axis is a log scale, and that we have a considerable amount of tracks of exactly 125, 256, and 512 frames because the views have been annotated for these number of frames.

Fig. 3 shows an example of a view from sample p02 with the annotated tracks for the last 25 frames overlaid. Notice that the example includes cells from all three motility categories. We observe this by visually inspecting the tracks and comparing them with the motility category definitions in Section 1.1.

[1] `https://chemometec.com/automated-cell-counters/xcyto-10/`

Table 1 Dataset sample information. Sample IDs are denoted as either donor (d) or pooled (p) samples, where donor samples are single anonymised sperm donor samples, and pooled samples are fully anonymised samples of raw sperm from three individuals pooled into one sample.

Sample ID	Est. conc. ($\cdot 10^6$ ml^{-1})	Sample handling	Views	#cells/view	Views with #frames annotated[1]: 125	256	512
d01	43.33	3x dilution	8	42.69	7	-	1
d05	-	Swim up	10	21.46	10	-	-
p02	-	-	4	101.77	3	-	1
p09	46.00	-	1	98.71	-	-	1
p15	17.38	-	8	47.65	7	1	-
p21	10.52	-	8	25.04	8	-	-
p22	55.21	-	2	121.98	2	-	-

[1] Views annotated for 125, 256, and 512 frames respectively

Table 2 Track ground truth information.

Frames	Independent samples	Videos	Tracks
125 frames	6	37	2,072
256 frames	1	1	73
512 frames	3	3	585
Total	7	41	2,730

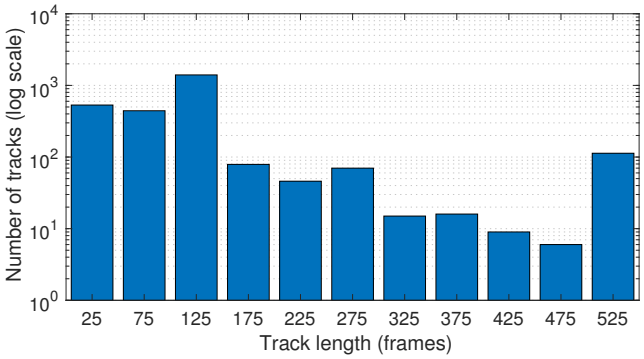


Fig. 2 Histogram of track lengths in our dataset.

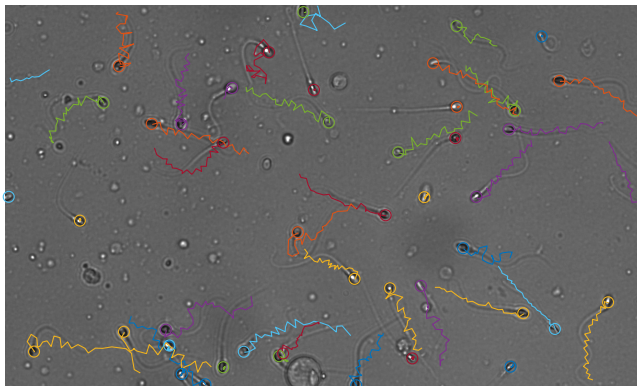


Fig. 3 Example cutout from a view from sample p02 with tracks for the last 25 frames overlaid.

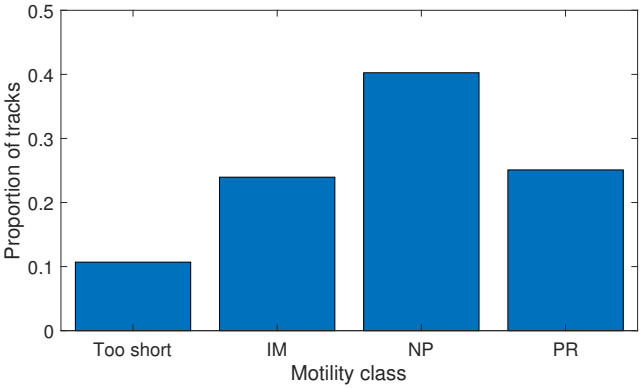


Fig. 4 Histogram of the track motility distribution of the dataset. The linearity of each track is measured continuously using a sliding observation window of 25 frames. The motility for each track is computed as the median of linearities measured.

2.2 Motility estimation

As mentioned in the introduction, we focus our investigation of motility using linearity (LIN), which is based on straight line velocity (VSL) and curvilinear velocity (VCL). VCL is defined as the velocity along the curvilinear track, and VSL is defined as the distance between the first and last point of a track divided by the duration. Finally, LIN is defined as the ratio between VSL and VCL. Formally, let $\boldsymbol{x}$ be a track of n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^2$. VCL, VSL, and LIN are defined as:

$$\mathrm{VCL}(\boldsymbol{x}) = \frac{f}{n-1} \sum_{t=2}^{n} \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_2 \tag{1}$$

$$\mathrm{VSL}(\boldsymbol{x}) = \frac{f}{n-1} \|\boldsymbol{x}_n - \boldsymbol{x}_1\|_2 \tag{2}$$

$$\mathrm{LIN}(\boldsymbol{x}) = \frac{\mathrm{VSL}(\boldsymbol{x})}{\mathrm{VCL}(\boldsymbol{x})} \tag{3}$$

where f is the fps and $\|\cdot\|_2$ is the euclidean norm.

Tomlinson et al. [18] base their motility estimates on VSL thresholds (a. $>$ 25 µm/s, b. 5-25 µm/s, c. $<$

5 µm/s, d. static) due to the use of the pre-2010 motility classes a-d are defined from speed rather than linearity of movement. Kraemer et al. [10] investigate various aspects of preparation handling and how the aspects affect the sperm kinematic characteristics measured. During their investigations they used the following LINintervals when describing the LIN distribution changes: $[0; 0.2[$, $[0.2 - 0.6]$, $]0.6; 1]$. Hidayatullah et al. [7] directly adopt these thresholds for categorisation into the three motility classes: IM ($[0; 0.2[$), NP ($[0.2; 0.6]$), and PR ($]0.6; 1]$). Their results show some correspondence with manual motility estimation, but the choice of thresholds have not been further investigated or validated. Fig. 4 shows a histogram of the track motility distribution of the entire dataset based on the manual annotations. LIN is measured continuously for each track using a sliding window of 25 frames, the median LIN is used as the track linearity, and the thresholds described above are used for categorising the track motility. Notice that approximately 10 % of the tracks are shorter than 25 frames. These tracks are typically located close to the edges of the view and therefore exit the view quickly.

2.3 Statistics

Throughout our experiments we use LIN distributions for our investigations. We base our experiments and results on non-parametric statistics in the form of box plots, scatter plots and Bland-Altman plots [1].

We use the scatter plots to visualise the correlation between different methods of measuring the linearity of an object. The two methods being compared are both error-prone, and thus we have to use a model that assumes error in both variables when estimating the linear correlation between the two methods. Given that our measures are equally scaled, we use the major axis method [21] for least squares fitting of the correlation plots.

Bland-Altman plots visualise the difference between two methods as a function of the mean of the two methods. Using these plots we measure the limits of agreement and the bias between two methods. We define the limits of agreement (LoA) as $\pm 1.45 \times$ Interquantile Range of the difference between the two methods (5% confidence interval if the differences are normally distributed), and the bias is the median of the difference between the two methods.

2.4 Motility estimation in practice

When conducting automatic motility estimation in practice we need to detect and track the sperm cells, after which the resulting track can be analysed. We use the method proposed by Nissen et al. [13] for detecting sperm cells. The method applies a convolutional neural network to a greyscale bright field image of a semen sample, computes connected components, and thresholds the components based on their area. The component centroids are used as cell locations. We use the recommended network called 3-conv-full-up-inc. The method is applied to each image of a video sequence giving us detections at every frame of the sequence.

The detections are linked together using the Noniterative Greedy Algorithm for Multiframe Point Correspondence (NGA) with mixed motion (linear and random) described by Smal et al. [17,16,9]. In short, the linker sequentially combines tracks frame by frame by minimising the cost of connecting previous and current detections with the detections in the next frame. The number of previous frames that are considered is determined by the buffer size Δt. A greedy choice of least-cost assignments is made based on costs calculated as a weighted sum of linear and random motion. The model requires tuning of a set of parameters, which we have manually optimised based on visual inspection of the resulting tracks (chosen parameters in parentheses): diffusion gate (80 pixels), directed motion gate (50 pixels), buffer size (5 frames), minimum track length (3 frames), initial speed of new tracks (8 pixels/frame), cost of missing detection (8), cost of track death (60), and cost of track birth (13). Notice that we operate with a 5 frame buffer. This makes the linker able to link tracks having at most three consecutive missing detections. The chosen detector only detects one object when multiple sperm cell head perimeters coincide or overlap, thus leaving only one detection for multiple tracks upon collision. These situations require a frame buffer big enough for the cells to spread apart and form individual detections. Choosing too large a frame buffer comes at the cost of computation time and typically decreases linking performance for cells having fewer consecutive missing detections.

We now describe how to quantify tracking performance to validate the detector and tracker described above. Given two sets of tracks (ground truth and linked tracks) we wish to pair the tracks such that we obtain the best possible total track overlap. Performance is computed on the basis of the tracks and their pairing. We following the instructions of Chenouard et al. [2] for pairing tracks using the Optimal Subpattern Assignment (OSPA) [15] method and report the tracking

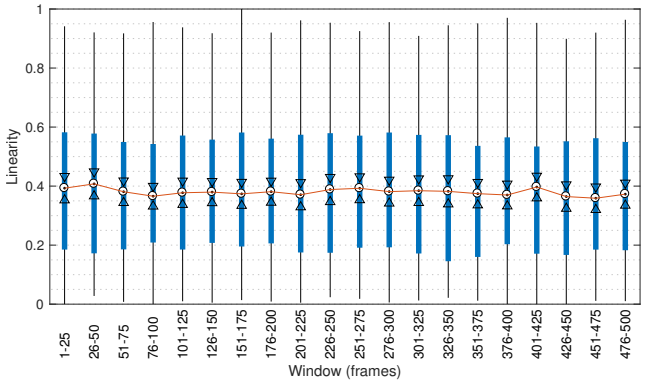


Fig. 5 Box plot of linearity measured for every second of the videos annotated for 512 frames. Window size is fixed to 25 frames (one second). For every window, only tracks that are annotated for the entire window are included.

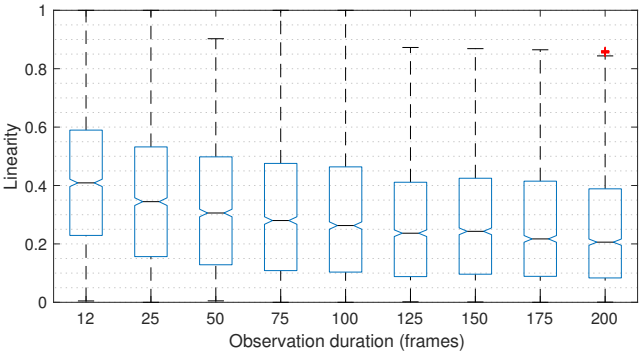


Fig. 6 Box plot of linearity for varying observation durations. The data has been sampled from tracks between 200 and 512 frames long (275 tracks). 2,000 tracks have been sampled for each observation duration. A random subtrack of the given observation duration was randomly chosen for linearity computation of each randomly selected track.

performance β with $\epsilon = 25$ described by Chenouard et al. [2]. Briefly explained, β gives a score between 0 and 1 based on the distance between paired ground truth and predicted tracks and penalties for non-paired tracks. The parameter ϵ defines a maximum distance penalty between track points.

3 Experiments and results

We here describe the experiments and results obtained in the investigation of each of the topics introduced in Section 1.2.

3.1 Metric behaviour

The first topic of our investigation is the metric behaviour. Generally the metric is not affected by the spatial position of the cell and therefore we focus on the temporal dimension.

3.1.1 Global motility drift

We hypothesise that the global drift in motility due to cell exhaustion cannot be observed in videos of 20 seconds. We investigate the hypothesis by selecting the videos annotated for 512 seconds, splitting them into 25-frame intervals, and computing the linearity distribution of all cells fully tracked within each interval. Fig. 5 shows the box plots of linearity for each of the intervals. The box plots show no general drift in the linearity distribution throughout the videos. Having confirmed this hypothesis, we can ignore the temporal position of tracks in the remaining experiments.

3.1.2 Linearity decrease

Our next hypothesis is that linearity decreases with increasing observation duration. In order to test this hypothesis we measure the linearity of a set of tracks for varying observation durations l. We test durations of 12, 25, 50, 75, 100, 125, 150, 175, and 200 frames (approximately 0.5 to 8 seconds). We only included tracks in the experiment which were at least as long as the longest duration to avoid introducing other biases in the experiment. Thus we only included tracks of at least 200 frames (275 tracks). Sperm cells may perform irregular behaviour and thus we sampled multiple random subtracks from each track to obtain a representative linearity distribution independent of track length and observation consistency. We conduct the following sampling for each observation duration l: First, we sample 2,000 random track indices. Second, we select a random subtrack of l points from each of the tracks identified by the previously sampled track indices. All random sampling was done using a uniform random distribution.

The results of the experiment are shown in Fig. 6. The figure shows a box plot of the sampled linearity distributions for the previously defined observation durations. Notice how the distributions are gradually pushed towards lower linearities as the observation duration increases. Table 3 shows motility classifications based on thresholding of the linearity distributions as explained in Section 2.2. Notice how the distribution changes from having 22.97 % PR cells with $l = 12$ to having only 9.40 % PR cells with $l = 200$. The experiment validates the hypothesis that linearity decreases for increasing observation duration, and thus the investigations of the remaining hypotheses need to account for this effect.

Table 3 Motility distributions for the data represented in Fig. 6 using the motility buckets based on LIN. IM: LIN $<$ 0.2, NP: 0.2 $\geq$ LIN $\leq$ 0.6, PR: LIN $>$ 0.6.

Obs. duration (frames)	IM	NP	PR
12	0.2212	0.5491	0.2297
25	0.3210	0.5020	0.1770
50	0.3673	0.4907	0.1420
75	0.3784	0.5095	0.1121
100	0.4141	0.4762	0.1097
125	0.4454	0.4654	0.0892
150	0.4383	0.4699	0.0918
175	0.4737	0.4256	0.1008
200	0.4910	0.4150	0.0940

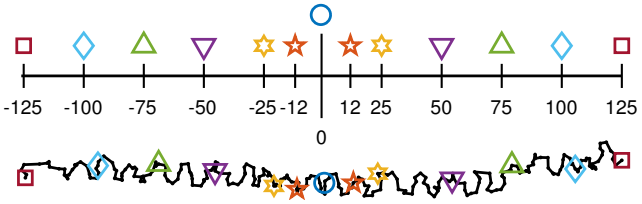


Fig. 7 Illustration of the temporal split of a track into two non-overlapping observations of varying sizes. The top line shows a temporal line offset/centred at the middle observation (frame 125) indicated by 0. Observation 1 is before and observation 2 is after the split. The bottom line shows an illustration of a track with the temporal markers defined in the top line at their spatial locations.

3.1.3 Metric consistency

The third hypothesis is that the consistency/stability of the linearity measured depends on the observation duration. We test this hypothesis by measuring the linearity of two non-overlapping observations of the same track and compare the two resulting linearities for each track. Given our data we look at observation durations l_1 and l_2 for observation 1 and 2 respectively of 12, 25, 50, 75, 100, and 125 frames (approximately 0.5 to 5 seconds). We find all tracks of at least 250 frames (196 tracks) and split them into two subtracks at frame 125. The subtrack before frame 125 is observation 1 and the subtrack after frame 125 is observation 2. Observation 1 always ends at frame 125 and observation 2 always starts after frame 125. Fig. 7 illustrates how a track is split temporally around frame 125 marked by 0 with markers for all observation durations on both sides of the split. The top line shows the temporal split and the bottom line shows a track with the corresponding temporal observation duration markers at their spatial locations along the track.

Fig. 8 shows scatter (top) and Bland-Altman (bottom) plots of the linearity LIN_{obs1} and LIN_{obs2} measured of observation 1 and 2 respectively for observation durations where $l_1 = l_2 = 12,\ 25,\ 50$, and 125. We have omitted the plots for durations 75 and 100 here for brevity, but they can be found in Appendix A. The identity line and linear least squared line found by using the major axis method [21] are also depicted in the correlation plots. We see that the data points get closer to the identity line as the observation duration increases from 12 (Fig. 8a) to 50 frames (Fig. 8c). The plots do not change much for longer observations. We also see that the least squared fitted line coincides well with the identity line indicating a linear correspondence between the linearities of observation 1 and 2. The Bland-Altman plots confirm these observations displaying decreased limits of agreement from 12 (Fig. 8e) to 50 frames (Fig. 8g) after which they stabilise.

Fig. 9 shows a plot of the median of difference and limits of agreement as a function of the observation duration from the Bland-Altman plots in Fig. 8. This plot verifies our observation of the limits of agreement being minimised and stabilised at observations of durations 50 to 125. We also see that the bias (median of the difference between the two observations) is an order of magnitude ($>$ 20 times) smaller than the limits of agreement and thus close to non-existent.

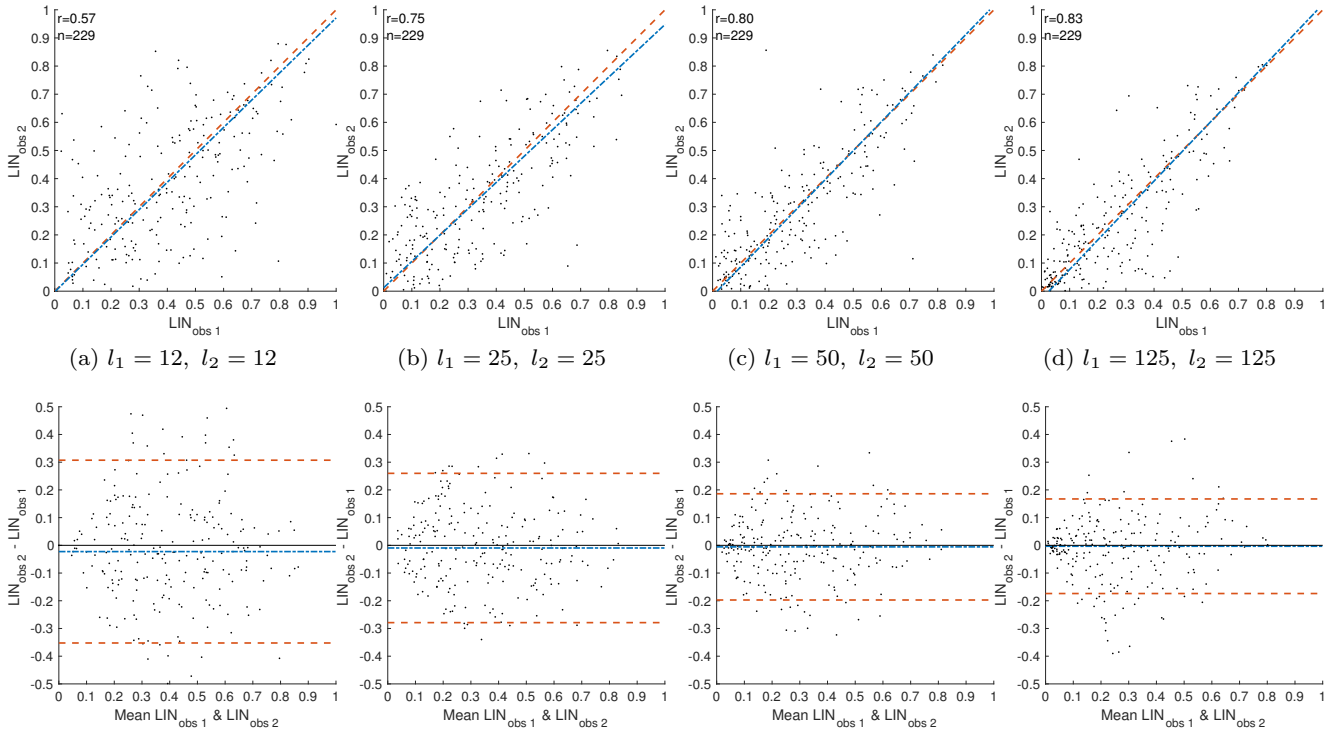
Given that an observation duration of 50 frames is the minimum duration for a stable linearity we re-do the experiment with $l_1 = 50$ and $l_2 = 12,\ 25,\ 50,\ 75,\ 100,\ 125$. Fig. 10 shows the bias (median of the difference) and limits of agreement for $l_1 = 50$ and varying observation 2 duration l_2. We notice that shorter l_2 durations display a tendency towards higher linearities and that longer l_2 durations display a tendency towards lower linearities as observed in Section 3.1.2. Also notice that the limits of agreement are stable for longer l_2 durations.

The results of the experiment validate the hypothesis that linearity consistency depends on observation duration. We have observed, that a duration of at least 50 frames optimises the linearity consistency.

3.2 Cell interaction

Cell interaction is the second topic of our investigation. We only have one hypothesis for this topic: motility is affected by cell collisions. In practice, collisions can be hard to link correctly, and they will in some cases cause tracks to die or spawn. We test this hypothesis in two parts: The effect of collisions on the linearity of individual cells, and the difference in linearity distributions of collision and non-collision observations. We test the latter in case the linker is unable to handle collisions.

For the first part, we find all tracks with at least one collision. The following is conducted for each ob-



(a) $l_1 = 12,\ l_2 = 12$ (b) $l_1 = 25,\ l_2 = 25$ (c) $l_1 = 50,\ l_2 = 50$ (d) $l_1 = 125,\ l_2 = 125$

(e) $l_1 = 12,\ l_2 = 12$ (f) $l_1 = 25,\ l_2 = 25$ (g) $l_1 = 50,\ l_2 = 50$ (h) $l_1 = 125,\ l_2 = 125$

Fig. 8 Correlation (top) and Bland-Altman plots (bottom) of LIN for two non-overlapping observations of the same track with varying observation lengths $l_1 = l_2$. In the top plots the dashed orange line indicates the identity line and the dot-dashed blue line indicates the linear least squares fitting. In the bottom plots the dashed orange lines indicate the limits of agreement and the blue dot-dashed line indicates the bias (median of the linearity difference).

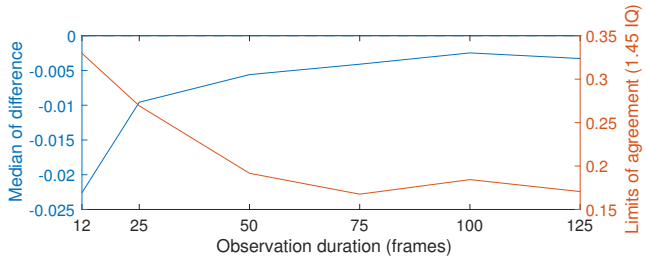


Fig. 9 Median of difference and limits of agreement for varying observation lengths where $l_1 = l_2$. The plot indicates consistency of linearity observations of various lengths.

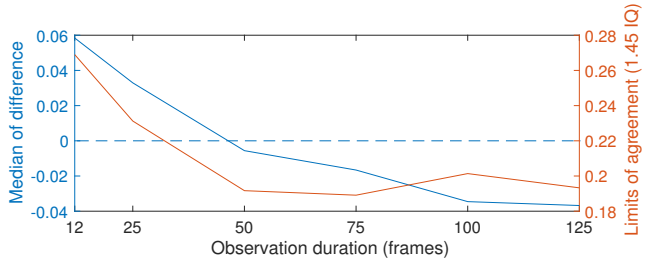


Fig. 10 Median of difference and limits of agreement for $l_1 = 50$ and varying l_2. The plot indicates bias.

servation duration $l = 25,\ 50,\ 75,\ 100,\ 125$: For each track we scan through all subtracks of length l and split them into two groups: subtracks with and without collisions. For each of the subtracks with collisions we count the number of collisions. We identify the subtrack with the maximum number of collisions having a non-overlapping subtrack from the non-collision group. We have conducted the Bland-Altman plot analysis for non-overlapping collision LIN_{col} and non-collision LIN_{nocol} observations of tracks for varying observation duration l. The plots are omitted here but can be found in Appendix B. Instead, Fig. 11 shows a summary of the bias (median of linearity difference) and limits of agreement for said Bland-Altman plots. We observe that the median of $\text{LIN}_{nocol} - \text{LIN}_{col}$ is very close to zero (-0.011 to 0.0069), and that the limits of agreement stabilise at 50 to 125 frames (0.12 to 0.10). This follows the observations in Section 3.1.3. The linearity of the individual cell is in other words not affected by collisions.

For the second part, we re-use the sampled data from Section 3.1.2 illustrated in Fig. 6. We split the data into two subsets: subtracks/observations with and without collisions occurring. Fig. 12 shows box plots for varying observation duration for non-collision and collision observations. We observe that the linearity distribution of non-collision observations i biased towards

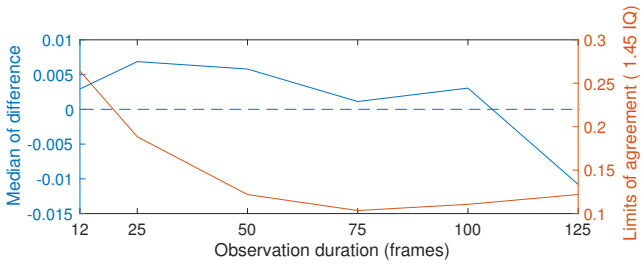


Fig. 11 Median of difference and limits of agreement for collisions and non-collision observations of varying observation lengths where $l_1 = l_2$.

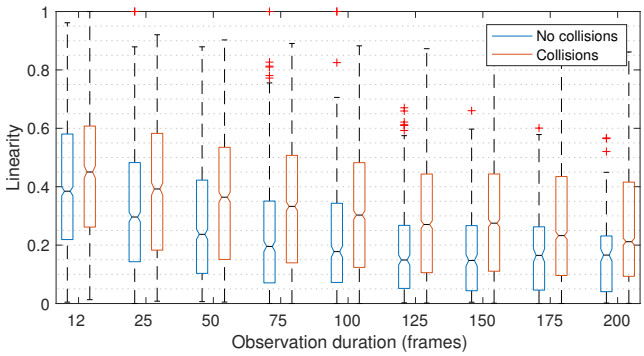


Fig. 12 Box plot of linearity for varying window sizes. The tracks are the same as in Fig. 6 but split based on whether a collision occurs within the sampled observation window. The blue plots show track samples with no collisions and the orange plots show track samples with collisions.

lower linearities than the distribution for collision observations for all observation durations. Notice that the box plots for long non-collision observations are based on a much smaller amount of observations because cells rarely are tracked for long durations without any collisions.

3.3 Spatio-temporal cell selection

Spatio-temporal cell selection is the third and last topic we investigate.

3.3.1 Cell origin

The first hypothesis of the topic states that cells entering the view during the analysis are biased towards progressively motile compared to origin cells. In order to test this hypothesis we compare linearity for fully tracked (origin) cells (start frame 1) and introduced cells (start frame > 1). We re-use the sampled linearity data from Section 3.1.2 and split it up into origin and introduced tracks. The tracks from which we sampled consist of 191 origin and 84 introduced cells. Given the uniform random sampling, we maintain the ratio between origin and introduced tracks for the sampled sub-

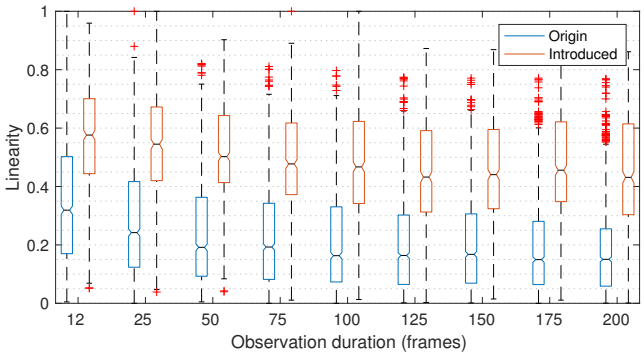


Fig. 13 Box plot of linearity for varying observation durations. The tracks are the same as in Fig. 6 but split based on start frame. The blue plots show origin tracks (start frame $= 1$) and the orange plots show introduced tracks (start frame > 1).

tracks. Fig. 13 shows two sets of box plots for varying observation durations for origin and introduced tracks. The distribution of introduced tracks is clearly biased towards higher linearities than the distribution of origin tracks. This holds for all observation durations. In other words cells entering the view during the analysis skew the distribution towards progressively motile independent of observation duration.

3.3.2 Field of view

The second hypothesis of the topic is that requirements on the track length cause a selection bias towards slow and immotile sperm cells. The reasoning behind this hypothesis is the difference in movement and the limited field of view (FoV). We investigate the hypothesis in two experiments: First, we look at the linearity distributions for different track lengths, and second, we investigate limiting the FoV based on the motion of moving cells.

In the first experiment we use the views annotated for the full 512 frames. From these views we select all tracks of at least 25 frames (511 tracks). We divide the tracks into subsets based on their track length and sample 2,000 random track indices from each subset. LIN is computed for a random subtrack of each of the randomly chosen track indices. Fig. 14 shows box plots of the linearity sampled for varying track length intervals between 0 and 512. We see that the maximum LIN decreases as the track length increases, and that cells of maximum length have a distribution significantly skewed towards lower LIN value than all the remaining track length intervals.

In the second experiment we investigate the likelihood that a moving cell stays inside the view within a specified duration by the use of a simulation of moving cells. The idea is to exclude cells too close to the

spatial edges of the view such that we keep a certain percentage of the moving cells within our observation boundaries after a specified duration.

We assume that cells are positioned at uniformly random positions inside the view, and that they move linearly in a uniformly random direction. The speed of cells vary greatly, and thus we choose to sample the speed of each simulated cell from an estimated linear speed distribution of our data. Sperm cells typically move in zig-zag motions making the average path somewhat linear. For each cell in our dataset we compute the average path using a sliding window of nine frames and use the median straight line speed of said average path as an approximation of the linear speed of the cell. Fig. 15 shows a histogram of the estimated linear speed distribution of moving cells (median VSL $> 10\,\mu m/s$ for the avg. path) which we sample the cell velocities from. We sample 500,000 cells and record their predicted positions after 25 and 50 frames.

We vary the FoV exclusion/inclusion distance d to the nearest edge between 0 and 165 µm. For each d we register the moving cells inside the FoV at the simulation start and the percentage of these cells still within the view at the simulation endpoint. We conduct this investigation for both the 25 and 50 frame simulations. Fig. 16 shows the fraction of cells inside the view and the distance from the view edge as a function of the FoV fraction of the view area included. For example, requiring 95 % of the moving cells to be within the view lets us utilise 76.20 % of the area for counting after 25 frames, which corresponds to a FoV distance of $23.69\,\mu m = 104.42$ pixels from the edge of the view. Using the same requirement for 50 frames means that we can only use 31.49 % of the view corresponding to a FoV distance of $81.28\,\mu m = 358.27$ pixels from the edge of the view.

3.4 Theoretical and practical example

The hypotheses presented in this work are all explored on the basis of ground truth tracks. In this experiment we investigate the difference between the theoretical and practical track linearity distribution. First, we verify our notion that linking performance decreases with increasing video duration.

To investigate how performance changes based on video duration, we choose the three videos annotated for 512 frames. The ground truth cell positions are linked using the NGA linker described in Section 2.4 for varying video durations from 12 to 500 frames, and the performance metric β is computed. Fig. 17 shows mean and standard error of β as a function of the video

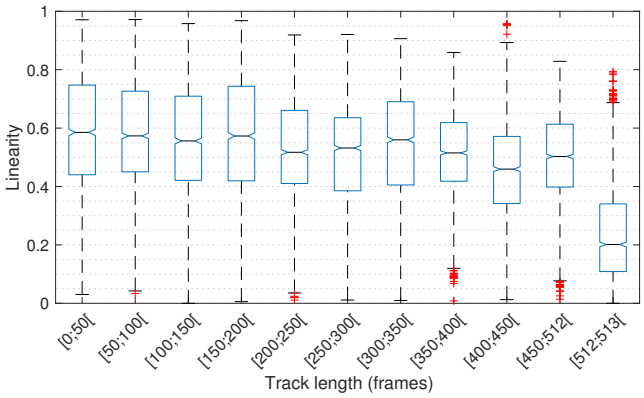


Fig. 14 Box plot of linearity for varying track length intervals. The data has been sampled from all tracks at least 25 frames long. 2,000 tracks have been sampled in each interval and for each a random subtrack of 25 frames was used for linearity estimation. Notice how the last box plot only contains cells annotated for the maximum 512 frames.

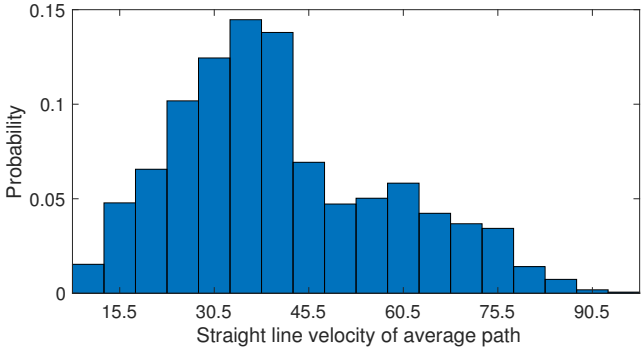


Fig. 15 Histogram of median straight line speed of the average path (9 point average) for all tracks. Tracks with a median speed lower than $10\mu m$ have been pruned to remove stationary cells.

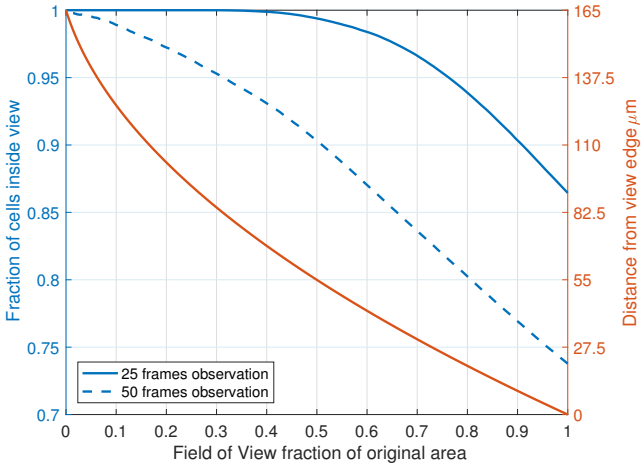


Fig. 16 Fraction of cells inside the full view (blue lines) and distance from view edge (orange) as a function of the FoV fraction of the full view area. The data is constructed from a simulation of 500,000 cells starting in random (uniform) positions inside the view, moving in random (uniform) directions with a random speed sampled from the discrete distribution in Fig. 15.

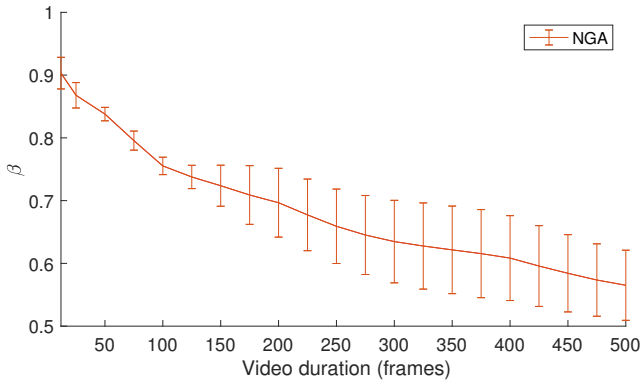


Fig. 17 Performance (β) as a function of the video duration.

duration. We see that performance decreases with increasing video duration as expected.

Second, we compare the linearity distributions from the ground truth and automatically obtained (practical) tracks. The experiment is conducted on all annotated videos and for both the first 25 and 50 frames for comparison. We obtain the practical tracks of each video by automatically detecting and linking sperm cells as described in Section 2.4. The tracks are analysed as follows: First we remove all cells starting outside the FoV define to keep 95% of the cells inside the view (Section 3.3.2). Second, we remove tracks shorter than the specific video duration. Third, we compute the linearity of the remaining tracks.

The experiment results in 192 boxplots which are found in Appendix C. For simplicity we compare the linearity distributions by comparing their medians. Fig. 18 shows the comparison of medians of linearity distributions. The comparisons are made by looking at histograms of the differences between linearity medians: Fig. 18a and 18b compare theoretical and practical medians on the view and sample level respectively, and Fig. 18c and 18d compare 25- and 50-frame distributions on the view and sample level respectively. Firstly we observe that the difference between linearity medians of theoretical and practical distributions are very close to zero. More than 80% of the views have a difference between -0.01 and 0.01 (Fig. 18a) and all the samples have a difference between -0.03 and 0.01 (Fig. 18b). These differences are in general slightly bigger for 50- than 25-frame tracks. Secondly, we observe the notable difference between 25- and 50-frame medians for both theoretical and practical tracks at view level (Fig. 18c). There is a slight tendency towards higher linearities for 25-frame medians skewing the histogram towards the positive side. This tendency is clearly observed when looking at sample level (Fig. 18d).

4 Discussion

4.1 Metric behaviour

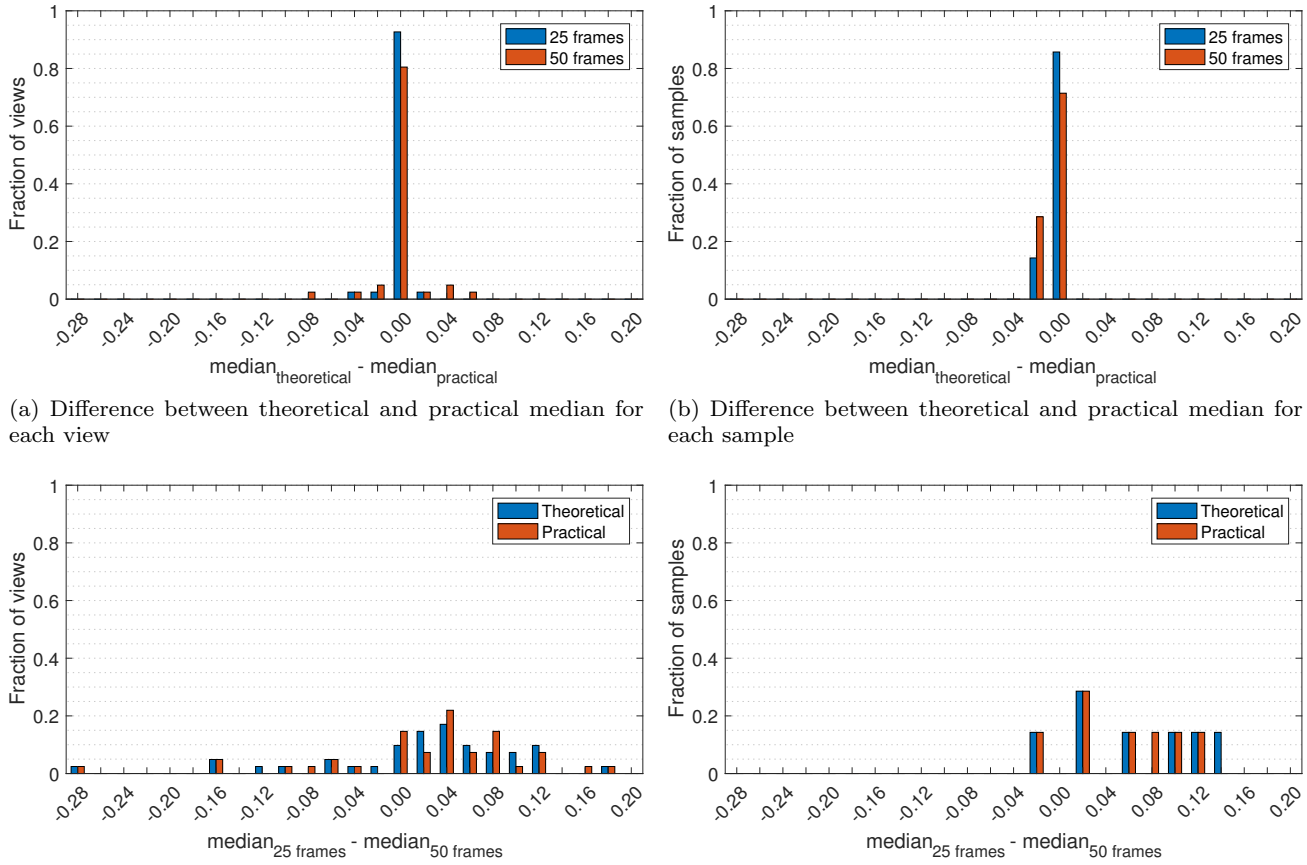
The experiments and results related to metric behaviour are all focused on investigating how the temporal dimension affects the linearity distribution measured. First, we briefly argue why we do not have a problem with global motility drift in our videos. Second, we discuss the observation duration and the two hypotheses related to the problem: linearity decrease and metric consistency.

4.1.1 Global drift

Urbano et al. [19] show that the track statistics for single tracks can change over longer periods. They presented specific data for two cells indicating large movement variations in at least one of the tracks slowing the movement down within the 55 seconds it was recorded. Motility is known to be affected by factors such as time passed since ejaculation and the sample storage temperature. As described in the results section 3.1.1 we do not observe any general drift in the linearity distributions of our 20 second videos. Though Urbano et al. [19] observed a change on single cell level, we do not observe it as a general drift on the population level. From our experiment we can conclude, that the level of motility is maintained throughout the 20 seconds with no drift. In other words the effect of cell exhaustion and change in sample temperature during the 20 second video-acquisition cannot be observed in our setup.

4.1.2 Observation duration

The duration of observation has been defined to a fixed length in most CASA instruments and automatic motility studies [7,10,18,19], and the choice of length is typically chosen by the authors based on recommendations from the WHO manual. The WHO manual [20] specifies that each sperm should be observed for at least one second, but the work referenced does not mention the parameter directly [12]. Other sources mentioned in the WHO manual do conduct a limited investigation of the problem [4,11](NOTE: Add more references?). The choice of one second seems reasonable: Most cells can be tracked for one second without exiting the view apart from the cells very close to the border. One second seems to be enough to make a manual judgement of the motility. The overall motility drift is non-observable as concluded above. In practice the minimum observation duration is desired for fast and correct detection and



(a) Difference between theoretical and practical median for each view

(b) Difference between theoretical and practical median for each sample

(c) Difference between 25 and 50 frame median for each view

(d) Difference between 25 and 50 frame median for each sample

Fig. 18 Histograms for comparisons of medians of linearity distributions on a view and sample level.

linking. There are however still issues to address before concluding anything: linearity behaviour and consistency

According to the WHO manual progressively motile cells are cells that move linearly or in big circles whereas non-progressively motile cells include all other movement patterns without progression such as cells moving in small circles. The difference between small and big circles are not described further. The choice of linearity as metric for determining motility seems reasonable from the definition of motility according to the WHO manual.

The definition of linearity has a natural built-in tendency to decrease as we increase the observation duration. A track consisting of two points always has linearity equal to 1 since the definitions of VSL and VCLare similar in this case. For short observation durations there is a significant risk that the cell moves inconsistently thus giving us a very noisy estimate of the general motility of the sperm cell. As the observation duration increases, an otherwise linear cell will have a higher risk of making a slight turn either from colliding with other objects or from irregular behaviour thus decreasing the linearity. Interestingly, cells moving in circles will have a very low linearity if the observation time is exactly equal to the time it takes the cell to move one full circle. The observation duration is in other words very important for the distribution of linearities we can expect to observe.

We investigate the change in the linearity distribution based on various observation durations in hypothesis 2. The results verify our intuition that the distribution of linearity based on longer durations cause a skew towards lower linearities. The bias identified when comparing non-overlapping observations of 50 frames with observations of varying durations further validated the statement. This means that we should be careful when using linearity for motility classification. We need to use a fixed observation duration for all cells analysed as also recommended by Davis and Katz [4] and Mack et al. [11]. The cutoff thresholds between the three motility classes cannot blindly be adopted by previous methods. They need to be identified for the specific observation duration and frame rate in a clinical study. One focus

point in such a study could be the distinction between small and big circles and the corresponding linearities measured. In order to be able to conduct the study (for 25 fps) we would have to extend our dataset with expert classifications of every single sperm cell track.

Mack et al. [11] concluded, that VSL, and VCL stabilised after 5-7 frames at 30 fps for five specific cells selected to cover varying movement progression. LIN was not directly investigated in this context, but due to the definition based on VSL and VCL we can assume, that LIN likewise stabilises after 7 frames for the cells they investigated. We investigate the stability of linearity for observation durations between 12 and 125 frames (approximately 0.5 to 5 seconds) on a population of 275 tracks with varying motility between completely stationary and highly linear. Based on the Bland-Altman plots we identify observations of 2 seconds to give maximum observation consistency/stability at the minimum number of frames. In other words the certainty of the linearity being representative for the cell motion is maximised at a duration of 2 seconds, and the certainty does not increase by increasing the observation duration to 5 seconds. This finding conflicts with the current belief that one second is enough to obtain a correct motility estimate.

4.2 Cell interaction

The second topic of our investigation is cell interaction and how it affects motility. Our results show that collisions do not affect the linearity measured of the individual cell as indicated by the stability of the limits of agreement. This indicates that cells are able to pass each other in the third dimension instead of repeatedly colliding or getting entangled. In other words it validates that the glass chamber depth of 20 µm is sufficient depth for motility analysis of human sperm cells, validating the statement by the WHO manual [20, pp. 138]: "Disposable counting chambers, 20 µm deep, give reliable results".

The number of collisions occurring in a sample is highly dependent on the concentration and motility. Current CASA systems have bounds on the concentrations they are applicable to due to the risk of introducing tracking errors upon collisions. We investigate if there is a linearity distribution difference between collision and non-collision cells. Naturally we would think that completely stationary cells have a smaller risk of being included in a collision than motile cells due to the fact that some other cell has to collide into the stationary one. The motile cells on the other hand can collide with both stationary cells and other motile cells. For short observation durations there is a small difference in the linearity distributions of collision and non-collision observations. For longer observation durations we see a bigger skew towards lower linearities for non-collision cells indicating that our notion from the previous statement is valid. If cell collisions are not handled by the linker, the collision tracks risk being filtered away due to the strict track-length policy described in Section 4.1.2, and thus the distribution skew mentioned above will be relevant. To summarise, cell collisions need to be handled by the linker to avoid selection bias, though collisions do not affect the linearity obtained for the single cell.

4.3 Spatio-temporal cell selection

Our goal is to measure the global motility distribution of a semen sample from a small view hereof. Ideally, we would like to instantaneously estimate the motility in the view allowing no cells to move in or out of the view during the analysis. The nature of sperm cell movement however require a certain temporal duration as shown in previous experiments. Correct selection of cells included in the analysis is therefore essential for an unbiased motility estimate. We split the topic into two parts: cell origin and field of view investigating the spatio-temporal positions of cells.

4.3.1 Cell origin

It seems like a natural deduction that cells moving into the view during the analysis skew the distribution of motility towards more motile cells. Introduced cells potentially also include immotile cells being pushed into the view by other cells or by drift in the sample in case the chamber is given insufficient time to settle after loading. We investigate the difference in linearity distributions between origin and introduced cells. The resulting statistics support the view that the linearity distribution of introduced cells is biased towards a higher linearity than the distribution of origin cells, and thus we need to exclude introduced cells entering the view during the analysis. This conclusion follows the advice stated in the WHO manual for manual counting [20, pp. 23]: "... avoid counting both those present initially plus those that swim into the grid section during scoring, which would bias the result in favour of motile spermatozoa".

In practice, cells risk being introduced during the analysis due to linker-errors causing tracks to terminate and new tracks to spawn wrongfully. This can be avoided by disallowing tracks within the interior of the view to die or spawn. The specific way of modifying or

developing such linkers would have to be further investigated in order to avoid introducing linker errors.

4.3.2 Field of view

The second part of the investigation of the hypothesis deals with specifying a FoV to make sure all cells initially identified are within the view after a specified observation duration. The problem of cells exiting the view was observed by Mack et al. [4] for tracks of 5 and 15 frames at 30 fps. They did not propose a way of fixing the skew introduced by the problem other than to use as few frames as possible; 5 frames for cell concentration and 15 frames for motility estimation. By simulating the average movement of sperm cells in our data, we found, that we can utilise 76.20 % and 31.49 % of the view for durations of 1 and 2 seconds respectively while maintaining 95 % of the cells initially identified inside the view. This means that we have to analyse nearly 2.5 times as many views for 2 seconds than for 1 second in order to count the same number of sperm cells. Naturally, the FoV distances reported depend on the cell motility distribution used for the simulation, and thus the distances for samples with other cell distributions could differ from the FoV distances reported here. Alternatively we could have used the maximum cell speed observed in our dataset for every cell simulated. This would give us the worst case FoV distance, which would hugely restrict the FoV and thus the number of cells we are able to count in each view. Our approach is a compromise between no FoV restriction and the worst case FoV restriction.

4.4 Theoretical and practical example

In Section 3.4 we first verify the statement that sperm cell tracking performance decreases with increasing video duration. The decreasing performance is naturally caused by an increasing amount of collisions in each track due to longer tracks. This effect causes us to favour short video durations to avoid introducing more linker errors than necessary to obtain sufficient motility statistics for each video. This view is supported by the previous discussion on observation duration and FoV restriction. As discussed earlier we need a duration of at least 50 frames in order to get the most consistent linearity measures of single cells.

Second, we investigate the theoretical and practical linearity distributions for 25 and 50 frame videos on the view and sample level. The resulting distributions are remarkably similar for most of the views despite the fact, that both the detector and linker introduce deviations from the ground truths. At sample level we can barely see any difference between the theoretical and practical distributions. We observe that there could be noticeable differences when comparing the results of 25 and 50 frame videos at view level. This difference is caused by the FoV restriction on views with a low number of sperm cells causing the statistics to be based on very few cells. The linearity distributions for 25 and 50 frames are very similar at sample level though the distributions for 50 frame videos are pushed slightly further towards lower linearities as expected by the linearity analysis discussed earlier. In conclusion, we can expect the linearity distribution obtained in practice to reflect the ground truth linearity distribution, and 25 and 50 frame videos result in comparable linearity distributions given a sufficient amount of cells in each sample.

5 Conclusion

We investigated how the movement of objects can be traced and described best so as to enable their subsequent classification. The three main topics of our investigation were: metric behaviour, object interaction, and spatio-temporal object selection. These topics were investigated for the specific application of human sperm motility analysis using the linearity motility measure. Our conclusions and recommendations are as follows:

1. Pre-define the observation duration and validate the motility estimation for this specific observation duration
2. An observation duration of 2 seconds gives maximum measurement consistency at minimum observation duration
3. Only include origin tracks, disregard tracks entering the view during analysis
4. Restrict the field of view based on observation duration and desired percentage of moving cells within the view at the end of the observation (Fig. 16)
5. Cell collisions should be handled by the tracker and cannot be ignored
6. No global linearity drift can be identified in videos up to 20 seconds
7. Aggregate motility statistics across multiple views of the sample for better global motility estimation. One view is typically not enough.

We compared theoretical (from ground truths) and practical (from automatically detected and tracked sperm cells) motility estimate distributions and achieved very similar results, concluding that we can expect practical results to reflect the theoretically possible motility estimate on similar data.

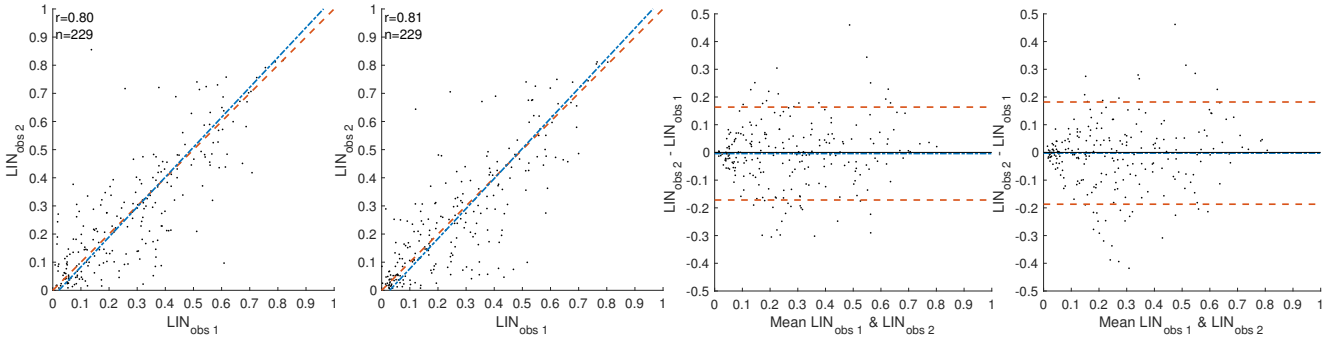
6 Conflicts of Interest

The manufacturer of the image cytometer used in our study was a partner in the jointly funded project. All experiments and evaluations of results were conducted without influence from any of the funding bodies. Søren Kjærulff works at ChemoMetec A/S but all studies have been performed independent of ChemoMetec A/S.

References

1. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet **327**(8476), 307 – 310 (1986). DOI https://doi.org/10.1016/S0140-6736(86)90837-8. URL http://www.sciencedirect.com/science/article/pii/S0140673686908378. Originally published as Volume 1, Issue 8476
2. Chenouard, N., Smal, I., de Chaumont, F., Maska, M., Sbalzarini, I.F., Gong, Y., Cardinale, J., Carthel, C., Coraluppi, S., Winter, M., Cohen, A.R., Godinez, W.J., Rohr, K., Kalaidzidis, Y., Liang, L., Duncan, J., Shen, H., Xu, Y., Magnusson, K.E.G., Jalden, J., Blau, H.M., Paul-Gilloteaux, P., Roudot, P., Kervrann, C., Waharte, F., Tinevez, J.Y., Shorte, S.L., Willemse, J., Celler, K., van Wezel, G.P., Dan, H.W., Tsai, Y.S., de Solorzano, C.O., Olivo-Marin, J.C., Meijering, E.: Objective comparison of particle tracking methods. Nat Meth **11**(3), 281–289 (2014). URL http://dx.doi.org/10.1038/nmeth.2808
3. Cooper, T.G., Yeung, C.H.: Computer-aided evaluation of assessment of "grade a" spermatozoa by experienced technicians. Fertility and Sterility **85**(1), 220–224 (2006). DOI https://doi.org/10.1016/j.fertnstert.2005.07.1286. URL http://www.sciencedirect.com/science/article/pii/S0015028205034308
4. Davis, R.O., Katz, D.F.: Standardization and comparability of casa instruments. Journal of Andrology **13**(1), 81–86 (1992). DOI 10.1002/j.1939-4640.1992.tb01632.x. URL http://dx.doi.org/10.1002/j.1939-4640.1992.tb01632.x
5. Egeberg, D., Kjærulff, S., Hansen, C., Petersen, J., Glensbjerg, M., Skakkebæk, N., Jørgensen, N., Almstrup, K.: Image cytometer method for automated assessment of human spermatozoa concentration. Andrology **1**(4), 615–623 (2013)
6. Gunnar, T., Anna, R.H., Ewa, T., Maryna, S., Aleksander, G.: Quality control workshops in standardization of sperm concentration and motility assessment in multicentre studies. International Journal of Andrology **28**(3), 144–149 (2005). DOI 10.1111/j.1365-2605.2005.00518.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2605.2005.00518.x
7. Hidayatullah, P., Awaludin, I., Kusumo, R.D., Nuriyadi, M.: Automatic sperm motility measurement. In: 2015 International Conference on Information Technology Systems and Innovation (ICITSI), pp. 1–5 (2015). DOI 10.1109/ICITSI.2015.7437674
8. van der Horst, G., Mortimer, S.T., Mortimer, D.: The future of computer-aided sperm analysis. Asian Journal of Andrology **17**(4), 545–553 (2015). DOI 10.4103/1008-682X.154312
9. Jaiswal, A., Godinez, W.J., Eils, R., Lehmann, M.J., Rohr, K.: Tracking virus particles in fluorescence microscopy images using multi-scale detection and multi-frame association. IEEE Transactions on Image Processing **24**(11), 4122–4136 (2015). DOI 10.1109/TIP.2015.2458174
10. Kraemer, M., Fillion, C., Martin-Pont, B., Auger, J.: Factors influencing human sperm kinematic measurements by the celltrak computer-assisted sperm analysis system. Human Reproduction **13**(3), 611–619 (1998)
11. Mack, S.O., Wolf, D.P., Tash, J.S.: Quantitation of specific parameters of motility in large numbers of human sperm by digital image processing. Biology of Reproduction **38**(2), 270–281 (1988). DOI 10.1095/biolreprod38.2.270. URL +http://dx.doi.org/10.1095/biolreprod38.2.270
12. Mortimer, D.: Laboratory standards in routine clinical andrology. Reproductive Medicine Review **3**(2), 97–111 (1994). DOI 10.1017/S0962279900000818
13. Nissen, M.S., Krause, O., Almstrup, K., Kjærulff, S., Nielsen, T.T., Nielsen, M.: Convolutional Neural Networks for Segmentation and Object Detection of Human Semen, pp. 397–406. Springer International Publishing, Cham (2017). DOI 10.1007/978-3-319-59126-1_33. URL http://dx.doi.org/10.1007/978-3-319-59126-1_33
14. Rehfeld, A., Dissing, S., Skakkebæk, N.E.: Chemical uv filters mimic the effect of progesterone on ca2+ signaling in human sperm cells. Endocrinology **157**(11), 4297–4308 (2016). DOI 10.1210/en.2016-1473. URL +http://dx.doi.org/10.1210/en.2016-1473
15. Schuhmacher, D., Vo, B.T., Vo, B.N.: A consistent metric for performance evaluation of multi-object filters. IEEE Transactions on Signal Processing **56**(8), 3447–3457 (2008)
16. Shafique, K., Shah, M.: A noniterative greedy algorithm for multiframe point correspondence. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(1), 51–65 (2005). DOI 10.1109/TPAMI.2005.1
17. Smal, I., Meijering, E.: Quantitative comparison of multiframe data association techniques for particle tracking in time-lapse fluorescence microscopy. Medical Image Analysis **24**(1), 163 – 189 (2015). DOI http://dx.doi.org/10.1016/j.media.2015.06.006. URL http://www.sciencedirect.com/science/article/pii/S1361841515000936
18. Tomlinson, M.J., Pooley, K., Simpson, T., Newton, T., Hopkisson, J., Jayaprakasan, K., Jayaprakasan, R., Naeem, A., Pridmore, T.: Validation of a novel computer-assisted sperm analysis (casa) system using multitarget-tracking algorithms. Fertility and sterility **93**(6), 1911–1920 (2010)
19. Urbano, L.F., Masson, P., VerMilyea, M., Kam, M.: Automatic tracking and motility analysis of human sperm in time-lapse images. IEEE Transactions on Medical Imaging **36**(3), 792–801 (2017). DOI 10.1109/TMI.2016.2630720
20. World Health Organization, et al.: WHO laboratory manual for the examination and processing of human semen (2010)
21. York, D.: Least-squares fitting of a straight line. Canadian Journal of Physics **44**(5), 1079–1086 (1966)

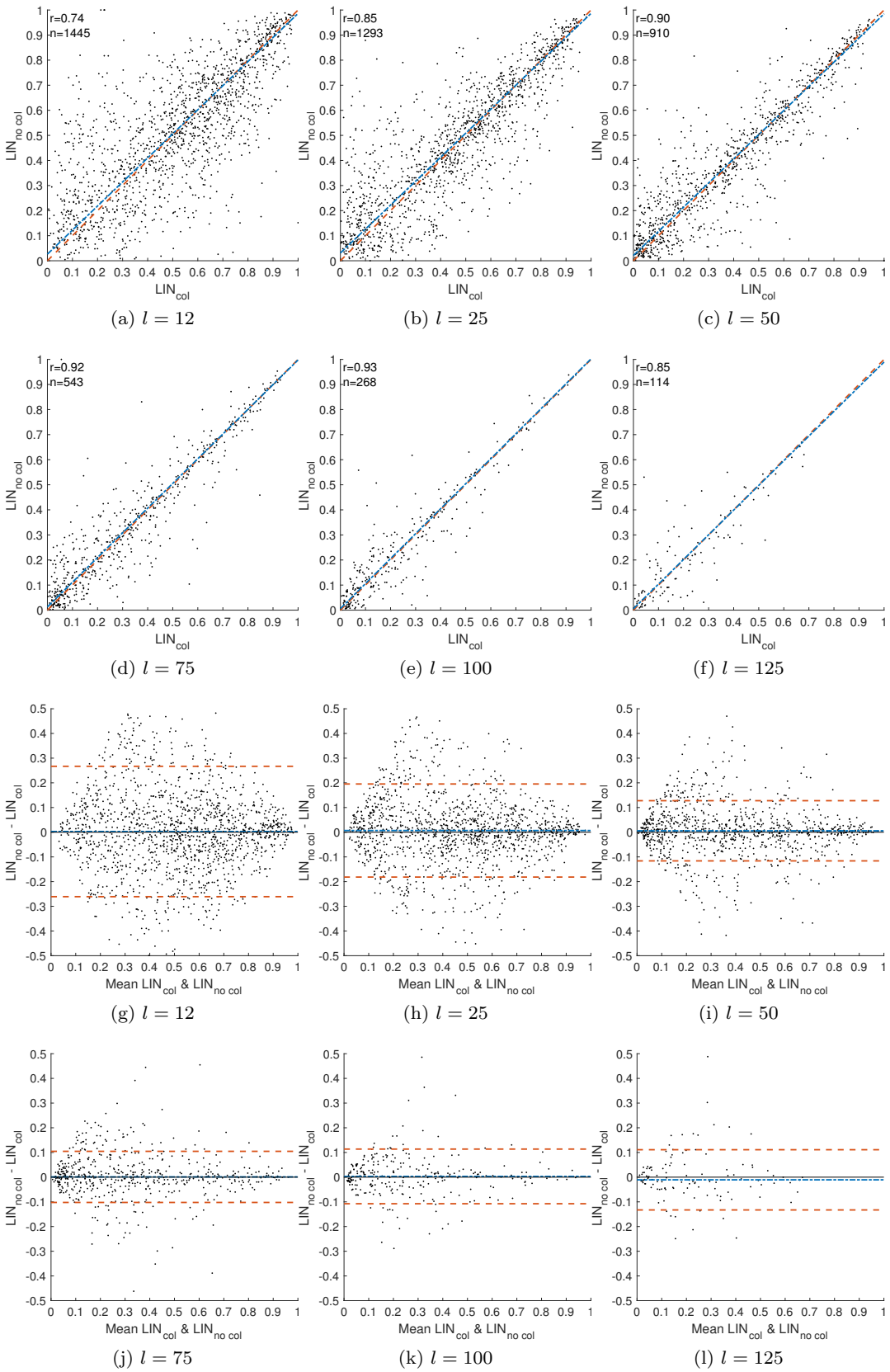
A Metric consistency



(a) $l_1 = 75,\ l_2 = 75$ (b) $l_1 = 100,\ l_2 = 100$ (c) $l = 75$ (d) $l = 100$

Fig. 19 Scatter (a-b) and Bland-Altman (c-d) plots of LIN for two non-overlapping observations of the same track with varying observation durations l_1 and l_2 respectively. In the scatter plots the dashed orange line indicates the identity line and the dot-dashed blue line indicates the linear least squares fitting. In the bottom plots the dashed orange lines indicate the limits of agreement and the blue dot-dashed line indicates the bias (median of the linearity difference).

B Cell interaction



(a) $l = 12$ (b) $l = 25$ (c) $l = 50$

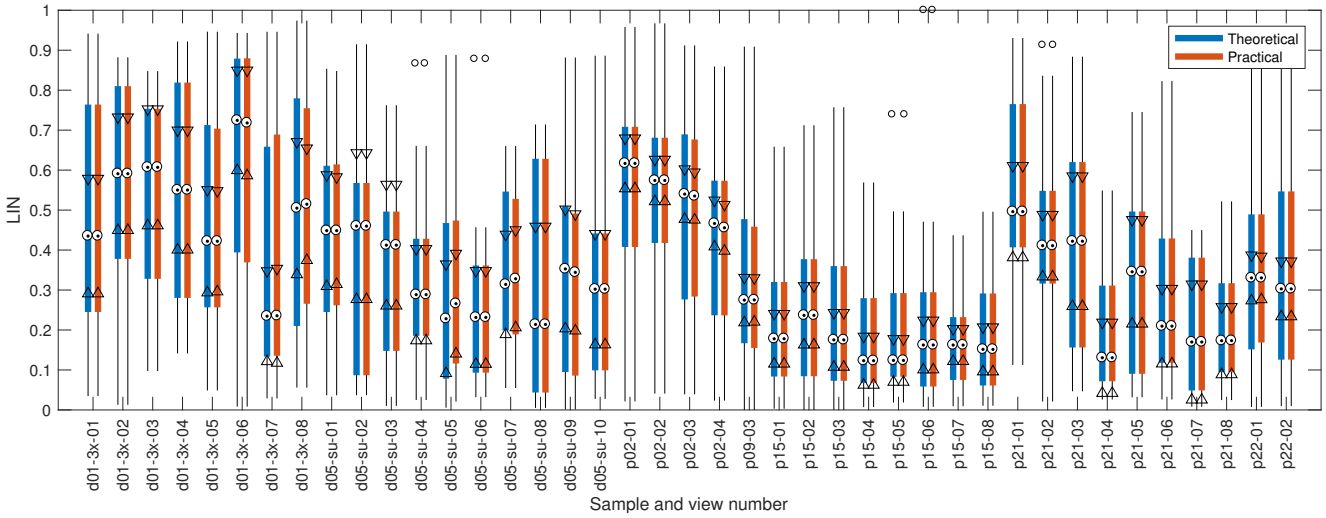
(d) $l = 75$ (e) $l = 100$ (f) $l = 125$

(g) $l = 12$ (h) $l = 25$ (i) $l = 50$

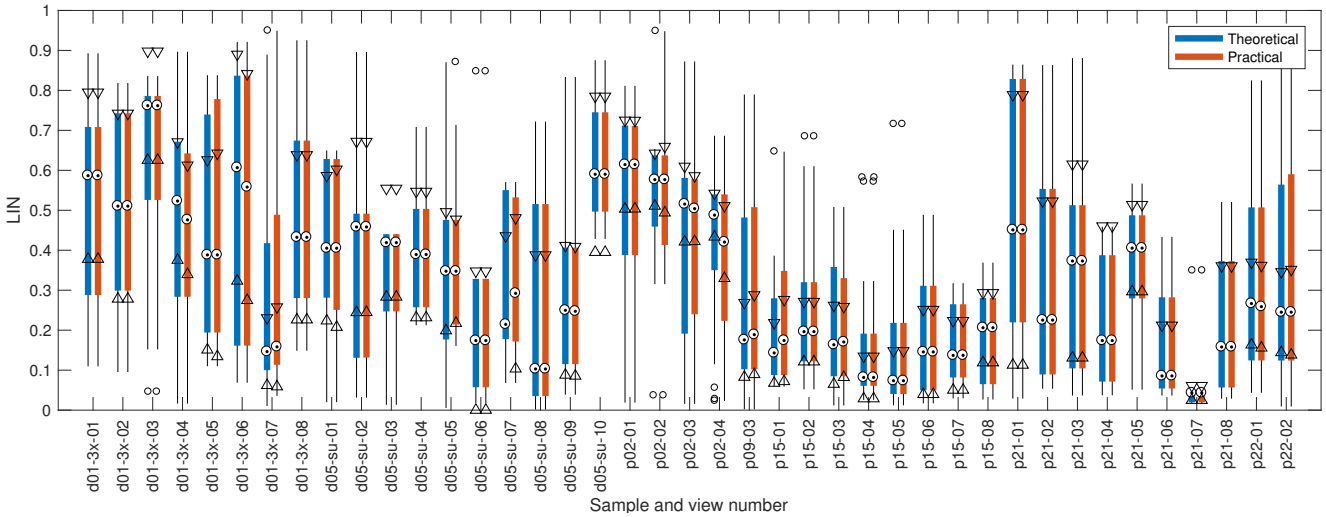
(j) $l = 75$ (k) $l = 100$ (l) $l = 125$

Fig. 20 Scatter (a-f) and Bland-Altman (g-l) plots of LIN for two non-overlapping observations of the same track where one observation contains collisions (LIN_{col}) and the other contains no collisions (LIN_{nocol}). The plots are made for varying observation duration l. In the scatter plots the dashed orange line indicates the identity line and the dot-dashed blue line indicates the linear least squares fitting. In the bottom plots the dashed orange lines indicate the limits of agreement and the blue dot-dashed line indicates the bias (median of the linearity difference).

C Theoretical and practical example

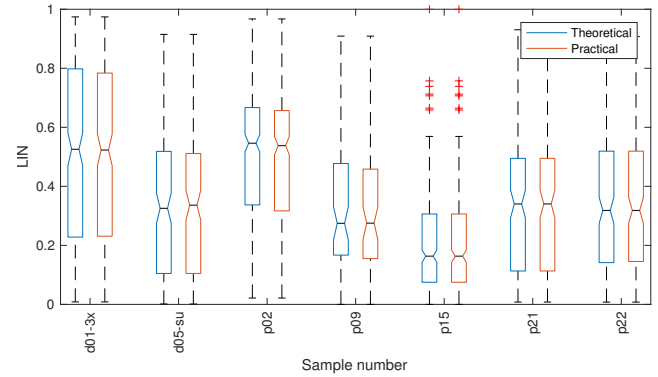


(a) $l = 25$ frames, $d = 23.81$ µm $= 104.01$ pixels

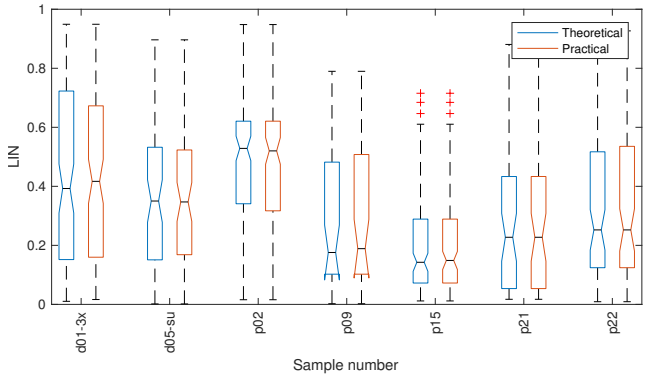


(b) $l = 50$ frames, $d = 81.98$ µm $= 358.12$ pixels

Fig. 21 Box plot of linearity distributions for all videos of l frames for the original ground truth tracks (theoretical, blue) and automatically linked detections (practical, red). Full length tracks are required, and tracks starting closer than d to the nearest edge have been removed.



(a) $l = 25$ frames, $d = 23.81$ µm $= 104.01$ pixels



(b) $l = 50$ frames, $d = 81.98$ µm $= 358.12$ pixels

Fig. 22 Box plot of linearity distributions for videos of varying frames l for the original ground truth tracks (theoretical, blue) and automatically linked detections (practical, red). Full length tracks are required, and tracks starting closer than d to the nearest edge have been removed. Tracks have been accumulated across all views from each sample.

Chapter 10

Paper III: Evaluation of a new integrated and fully automated system for sperm motility analysis

Andrology manuscript
(to be submitted)

Evaluation of a new integrated and fully automated system for sperm motility analysis

Malte Stær Nissen[1,2,3] · **Mads Nielsen**[1] · **Kristian Almstrup**[2,3]

Received: date / Accepted: date

Abstract

Background:
Sperm motility analysis traditionally reports motility grades (A-D) based on visual observations. Several computer-aided systems for sperm motility analysis exist but no unifying and platform-independent conversion exist that recapitulate the manual evaluation.

Objectives:
To evaluate a new integrated and fully automated system for sperm motility analysis and to identify conversion parameters recapitulating manual analysis.

Materials and methods:
Acquisition of motility tracks were facilitated by image cytometry and analysed for durations of 1 and 2s. Manual and automated motility data were acquired from 77 ejaculates. Straight line velocity (VSL) was used for conversion between kinematic parameters and motility grading. The VSL grading thresholds were estimated achieving minimal bias between manual and automatic read-outs followed by an analysis of temporal, intra-aliquot, and inter-aliquot variation of the automated read-outs.

Results: Using WHO thresholds we observed a significant bias (1s: bias -14, p= 0.01; 2s: bias -19, p=001) between manual and automatic read-outs of AB%. Optimising the thresholds to reflect manual read-outs yielded a good correlation with non-significant bias for both AB% and D%. Temporal variation of automatic AB% was minimal but linearity was significantly affected. Intra-aliquot variation was slightly higher than temporal (1-2s: 14 vs. 1s: 4.7, 2s: 3.9) and inter-aliquot variation was considerable (1s: 31, 2s: 26). Automatic analysis performed slightly better than the manual when forcing read-outs to obey the Poisson distribution as recommended by the WHO.

Discussion:
After optimising the gating thresholds, we observed minimal bias between the manual and the automated analysis. However, the intrinsic variation was still considerable. The automated system performed slightly better than manual analysis and allow technicians to focus more on sample handling than performing the counting itself.

Conclusion:
By optimising the VSL-gating thresholds we achieved comparable results between the automated and manual methods. Linearity was more affected by temporal variation than AB%.

Keywords CASA · Automatic motility estimation · Motility grading

This work is partly funded by the Innovation Fund Denmark (IFD) under File No. 4135-00169B, and by ChemoMetec A/S, Gydevang 43, DK-3450 Allerod, Denmark

M. Nissen
E-mail: nissen@di.ku.dk
[1]Dept. of Computer Science, University of Copenhagen, Copenhagen, Denmark
[2]Dept. of Growth and Reproduction, Rigshospitalet, University of Copenhagen, Denmark
[3]International Center for Research and Research Training in Endocrine Disruption of Male Reproduction and Child Health (EDMaRC), Rigshospitalet, University of Copenhagen, Denmark

1 Introduction

Today, routine analysis of human semen samples is conducted manually which is very labour intensive. The examination and analysis of human semen was first standardised by the World Health Organisation (WHO) in 1980 [1], and the 5th and the latest edition of the standard was published in 2010 [15]. The analysis includes estimating parameters such as sperm concentration, motility, morphology, and vitality. In this study, we focus on human sperm motility analysis.

The recommendations/guidelines for sperm motility grading and analysis has been re-defined multiple times during the past 30 years. The WHO defined four motility grades (A-D) based on swimming speed in 1999 [14]. Cooper & Yeung [3], however, discovered that in practice the differentiation between the four motility grades was too difficult to estimate for technicians. To overcome this issue a new definition was introduced in 2010 having three motility grades based on human-interpretable descriptions [15]: Progressive motility (PR), non-progressive motility (NP), and immotility (IM) corresponding to grade AB, C, and D, respectively.

Since the 1980s many have tried to autmate motility analysis using Computer-Aided Sperm Analysis (CASA), but many problems still limit the clinical applicability of CASA in the routine analysis as described by van der Horst et al. [8]. Generally all CASA machines measure the same set of basic sperm kinematic metrics for each sperm cell described by the WHO [15, Section 3.5.2.3, pp.138-139]: curvilinear velocity (VCL), straight-line velocity (VSL), average path velocity (VAP), amplitude of lateral head displacement (ALH), linearity (LIN), wobble (WOB), straightness (STR), beat-cross frequency (ALH), and mean angular displacement (MAD).

While the WHO grades builds on practical grouping of what can be observed by the eye, the CASA systems measure different kinematic parameters. Currently, there exists no clearly defined and 100% accurate translation between manual motility grading and the automatically obtained motility metrics. Cooper & Yeung [3] based their motility grading on VSL before the re-definition of motility in 2010. Hidayatullah et al. [7] used a gating on LIN, and Goodson et al. [6] reported that proprietary systems currently use a mixture of gates on metrics such as VAP and STR to estimate the motility grading. Mortimer et al. [9] used the following metric gating for classifying "'good mucus-penetrating' kinematic characteristics": VAP $\geq 25\,\mu m$ and STR $\geq 80\,\%$ and ALH $\geq 2.5\,\mu m$ where ALH is the amplitude of lateral head displacement meaning the distance between the average path and the curvilinear path. Recently, Goodson et al. [6] developed the CASAnova algorithm for classifying human-like motility grades with high accuracy. However, several metrics used in this work depend on the frame rate [10], and thus the specific model parameters reported can not be applied to other experimental setups.

In this study we evaluate a new fully automated sperm motility analysis system, based on image cytometry, by comparing read-outs of manual and automatic motility analysis. We optimise the gating thresholds of the automatic analysis and thoroughly evaluate the degree of variation in read-outs of the automatic motility analysis.

2 Material and methods

2.1 Semen samples

We collected data for 77 semen samples with varying sperm motility from January to June 2018. All individuals were patients at the Department of Growth and Reproduction, Copenhagen University Hospital (GR). 10 of the patients were from infertile couples referred to GR for routine andrological workup and 67 of the individuals were referred from general practitioners for the initial routine assessment of semen quality. The samples were collected at the department by ejaculating into a (wide-mouthed) plastic container. The samples were then allowed to liquefy for at least 15 minutes in an incubator at 37 °C before being manually and automatically analysed as described in the sections below. The time between ejaculation and incubation was between 1 and 18 minutes with a median of 5 minutes. For this study we gathered the following information about each sample: time (min) between ejaculation and incubation, time (min) between ejaculation and manual analysis, time (min) between ejaculation and automatic analysis, manually counted motility numbers (AB, C, D) according to WHO guidelines (see further details below), automatically acquired videos of the sperm samples for further analysis (see below), sperm dilution estimated by the technician, and sperm concentration measured using the NC-3000^{TM} image cytometer (ChemoMetec) as described by Egeberg et al. [4,13].

2.2 Manual motility estimation

The manual motility estimations were conducted according to the guidelines described by the WHO [15]. Briefly described, two 10 µL aliquots of the semen sample were placed on a microscope slide (Menzel Gläser), and a $22 \times 22\,mm$ cover slip was placed on top of each aliquot creating volumes with a depth of approx. 20 µm [15, pp. 18, Section 2.4.2]. A phase-contrast microscope with a 20x objective, 40x objective, and a heated stage pre-heated to 37 °C was used to assess motility after the sample drift had settled. Firstly, the visual appearance between the two aliquots was checked for consistency, and the dilution of the sample was estimated using the 20x objective. In case of visual appearance inconsistency two new aliquots were taken and the entire process was repeated. Secondly, the sperm motility was estimated by counting sperm cells and grading them as progressively motile (AB), non-progressively motile (C), or immotile (D). 200 sperm cells were counted in each aliquot (a total of 2x200) and the counts were checked for consistency. The check consisted of computing the average frequency of each of the three motility grades. The average frequency of the most frequent class was chosen and the frequency difference between the two aliquots was tested based on Table A7.2 of the WHO manual [15, pp. 259]. This table is based on the discrete Poisson probability distribution. If the frequency difference was within the allowed margin, the average motility estimation of each grade was accepted as the estimated sperm motility of the semen sample.

Otherwise two new aliquots were taken and the procedure repeated. In our study only 2x100 cells were counted due to limitations on lab technician time, but the quality control criteria for having counted 2x200 cells were used. This choice meant that samples were counted faster, but the stricter limits of agreement between aliquot counts were maintained and thus the quality of the analysis was maintained. In theory, counting only 2x100 cells imply that more samples would have to be re-counted due to failing quality control. In practice, the approach was used at GR prior to our study with a very low re-counting frequency leading to a significantly shorter analysis time. The laboratory at GR has taken part in the ESHRE semen analysis quality control program for several years. This entails alignment of manual motility estimates across laboratories.

2.3 Automatic motility estimation

The automatic motility estimation was performed in two rounds with slight differences. Sample preparation differed slightly between the the two rounds of investigation whereas video acquisition and analysis was conducted the same way for all samples.

The first 53 samples were prepared as follows: A small aliquot (approx. 60 µL) of the semen sample was transferred to an adjacent lab for processing. The aliquots were kept at 37 °C for varying time periods before processing. For samples with an estimated dilution of 1:20 or lower, 10 µL of the raw semen was used. Samples with higher dilution estimate the raw semen was diluted using human tubluar fluid (HTF^+) medium as described by Egeberg Palme et al.[5] to achieve a dilution of approximately 1:20. Two times 10 µL were placed separately on a microscope slide (Menzel Gläser), covered by a 22×22 mm cover slip each and loaded into an Xcyto® 10 image cytometer (XC10, ChemoMetec) and video acquisition started within a few minutes.

The last 24 samples were prepared as follows: Two 10 µL aliquots were taken from each sample at the same time where aliquots for manual analysis were drawn and placed on a microscope slide and covered by a 22×22 mm cover slip avoiding air bubbles. The aliquots were examined for consistency and drift under a phase-contrast microscope with a 20x objective. If the aliquots drifted too much or the samples were too inconsistent, two new aliquots were taken and the process repeated. The aliquots were loaded into an XC10 and video acquisition was conducted once the observable drift had disappeared.

We conducted the video acquisition by capturing videos (1920×1440 pixels) of 4 seconds at a rate of 25 frames per second for up to 15 views of each aliquot. Each view spans an area of 435.6×326.7 µm. This process took approx. 5 minutes per aliquot. After acquisition we manually registered views containing air bubbles, big lumps of agglutinated sperm cells, or sample drift, and removed these from the subsequent analysis. The videos were digitally analysed as described in detail by Nissen et al. [12] for both 1- and 2-second analysis. In brief, the analysis of videos followed these steps: first, sperm cells are identified in each frame using image segmentation [11]. Second, the sperm cell locations are linked across frames to form tracks throughout the videos. Third, the tracks are filtered to achieve an unbiased estimate of the motility metrics. The outcome of the analysis is a list of sperm tracks for each of which the standard CASA metrics are computed as described by the WHO [15, Section 3.5.2.3, pp. 138-139]. We utilise the linearity (LIN) and straight line speed (VSL) metrics in our work.

The motility grading definition was changed from the 4th to the 5th version of the WHO manual as described in the introduction. The new definition is mathematically very vaguely formulated and follows the intuition of motility grading made by humans, whereas the old definition very precisely described the grading from a mathematical point of view. The motility grading of the 4th edition of the WHO manual [14, pp. 9-10] was defined as follows:

a) "rapid progressive motility (i.e., ≥ 25 µm/s at 37 °C and 20 µm/s at 20 °C; note that 25 µm is approximately equal to five head lengths or half a tail length);
b) slow or sluggish progressive motility;
c) nonprogressive motility (< 5 µm/s);
d) Immotility."

The sperm swimming speed is not further defined, but as mentioned earlier, VSL or VAP is often used. We chose to use VSL for motility grading since it is shown to be independent of frame rate [10] and is computed the same way across all CASA systems as opposed to VAP which fulfils neither of these properties. In our work we mainly focus on the percentage of progressively motile sperm, which we defined as the percentage of cells in categories a) and b) and abbreviated as motility grading AB.

The three motility grades progressive motility (AB), non-progressive motility (C) , and immotility (D) are separated by placing two gating thresholds on VSL: the AB threshold t_{AB} and the D threshold t_D. Tracks with VSL $\geq t_{AB}$ are classified as AB, tracks with $t_{AB} >$ VSL $\geq t_D$ are classified as C, and tracks with VSL $< t_D$ are classified as D. The AB threshold is defined as $t_{AB} = 5$ µm/s according to the motility definition shown above, whereas the D threshold is not defined.

2.4 Agreement between repeated measurements

The quality control conducted during manual motility analysis is described by the WHO [15]. It is based on a 95% confidence interval on the Poisson distribution. Specifically, the difference $d = |p_1 - p_2|$ between probability 1 p_1 and probability 2 p_2 has a limit of $1.96\sqrt{\frac{2\bar{p}(100-\bar{p})}{N}}$ where $\bar{p} = \frac{p_1+p_2}{2}$ and N is the number of objects counted in each aliquot. In the automated analysis, we set N to the average of the two counts in case of estimates based on different counts.

2.5 Comparison of manual and automatic motility

We compared the manual and automatic motility scores based on the average of the two aliquots of each sample. The comparison was conducted in two rounds. First, we compared the AB% based on the AB threshold defined by WHO: $t_{AB} = 5\,\mu m/s$ Second, we measured and minimised the bias between the manual and automatic AB% by varying the AB threshold and the bias between the manual and automatic D% by varying the D threshold. Using the new thresholds we compared the manual and new automatic AB% and D%. Only samples with less than an hour between the manual and automatic motility analysis and at least 100 tracked sperm cells in each aliquot were included in the threshold estimation.

Sperm sample motility declines over time, but the decline varies from sample to sample. We therefore split the comparisons into three parts depending on the time (hours) between the manual and automatic motility analysis Δt. Analyses with low Δt are expected to show a high AB% inter-analysis correlation, whereas analyses with high Δt are expected to reflect the motility decline variation.

2.6 Temporal variation

We estimated the temporal variation of the metrics measured by the automatic analysis within each aliquot by collecting the tracks from all views of the same aliquot and calculating the AB% for each aliquot at non-overlapping time steps for both 1 and 2 seconds.

2.7 Intra-aliquot variation

The intra-aliquot variation was estimated by splitting the views of each aliquot into two sets of views (view splits) with as equal a total track number as possible. The AB% of each set of views from the same aliquot was compared.

2.8 Inter-aliquot variation

To estimate the inter-aliquot variation we compared the AB% of aliquot 1 and 2 independently for automatic and manual motility analysis. For the automatic analysis we conducted the comparison for both the subset of samples that lives up to the quality control and for all samples.

2.9 Statistical analysis

We use Bland-Altman (BA) plots [2] to investigate the bias and limits of agreement (LoA)/coefficient of reproducibility between two measures of motility, and we use the Major axis regression (MA) [16] for estimating the linear relationship between the two measures. The resulting linear relationship equation $y = ax + b$, the Pearson r-value squared r^2, and the number of datapoints n included in the analysis is shown on each of the scatter plots. We define the coefficient of reproducibility as the 95% confidence interval for the difference between the two methods being compared. For normally distributed differences we define $\text{RPC} = 1.96\sigma$, where σ is the standard deviation of the differences, and for other distributions we use the non-parametric equivalent $\text{RPC}_{\text{np}} = 1.45\text{IQR}$, where IQR is the inter-quantile range $(Q_3 - Q_1)$ of the differences. All statistics and plots are created in MATLAB® 2016b[1].

3 Results

3.1 Comparison with manual motility

Manual counting is currently the method recommended by the WHO to assess motility. We therefore aimed to compare our newly developed method (Nissen et al. [12]) for motility assessment to manually determined measurements.

First, we compared the manual and automatic analysis using the WHO defined thresholds. Fig. 1 shows scatter plots for the comparison between manual and automatic motility analysis using the WHO defined AB threshold. The measurements we compared were averages of the two aliquots of each sample for both the manual and automatic methods. The plots are shown for both 1-second analysis (Fig. 1a-c) and 2-second analysis (Fig. 1d-f) and for the sample criteria $\Delta t \geq 1$ (Fig. 1a, 1d), $\Delta t > 1$ (Fig. 1b, 1e), and all Δt (Fig. 1c,

[1] MA code: `http://www3.mbari.org/Products/Matlab_shell_scripts/regress/lsqfitma.m`, BA code: `https://se.mathworks.com/matlabcentral/fileexchange/45049-bland-altman-and-correlation-plot`

1f). Only samples with automatic analysis of at least 100 sperm cell tracks in both aliquots (2x100) were included. We observed a linear correlation ($r^2 = 0.55$ and $r^2 = 0.31$ for 1- and 2-second analysis, respectively) with a few outliers for samples with $\Delta t \leq 1$ and with a bias towards lower AB% for automatic analysis. Samples with $\Delta t > 1$ scattered more towards lower AB% for the automatic analysis but with a high amount of observations within the same correlation range as for $\Delta t \leq 1$. These observations were similar for both 1- and 2-second analysis.

Fig. 2 shows BA plots of the data shown above with $\Delta t \leq 1$. The 1-second analysis from $n = 24$ samples has $\text{RPC}_{\text{np}} = 17$, and a bias of -13.7 with a p-value of 0.0122. The 2-second analysis from $n = 20$ samples has $\text{RPC}_{\text{np}} = 18$, and a bias of -19.0 with a p-value of 0.0011.

Second, we investigated the bias as a function of the AB and D threshold definitions. We minimised the bias and compared the resulting automatic method to manual analysis. Fig. 3 shows the bias and RPC_{np} as functions of the AB threshold (Fig. 3a) and D threshold (Fig. 3b). The figures show results for both 1- and 2-second analysis. Generally, the AB% bias decreased as the AB threshold increased (0-7) and the RPC_{np} decreased slightly from approx. 30 to 20. When varying the AB threshold the RPC_{np} was nearly identical for 1- and 2-second analysis whereas the bias was higher for 1- than 2-second analysis. When varying the D threshold both the D% bias and RPC_{np} were slightly lower for 1- than 2-second analysis. We identified the thresholds yielding non-significant biases closest to zero which are shown in table 1. Using these thresholds we computed the automatic AB% and compared against the manual method. These comparisons are shown in Fig. 4. In general we observed a correlation along the identity line with varying degree of scattering. After the optimisation we achieved higher RPC_{np} and slight changes of r^2-values in both directions. Samples with $\Delta t \leq 1$ in general showed smaller signs of scattering than samples with $\Delta t > 1$. Both the 1- and 2-second analysis had two "outlier" samples which were automatically analysed 0.4 and 0.57 hours (24 and 34 minutes) after the manual motility was estimated.

While the comparison between manual and automated analysis were affected by the chosen AB and D threshold we tried to compare only the fraction of immotile cells. Fig. 5 shows scatter plots of D% (immotile sperm cells) with the optimised D thresholds for 1- and 2-second analysis for all Δt. These plots indicate a linear correlation with a large amount of variation yielding $r^2 = 0.29$ and $r^2 = 0.21$ respectively with non-significant biases of -0.71 and -0.14.

Threshold	Info	25 frames	50 frames
AB	Threshold	1.80	1.10
	Bias	0.2	0.3
	Bias p	0.70	0.86
	RPC_{np}	22.8	25.7
D	Threshold	1.05	0.75
	Bias	0.6	0.3
	Bias p	0.99	0.90
	RPC_{np}	26.1	27.2

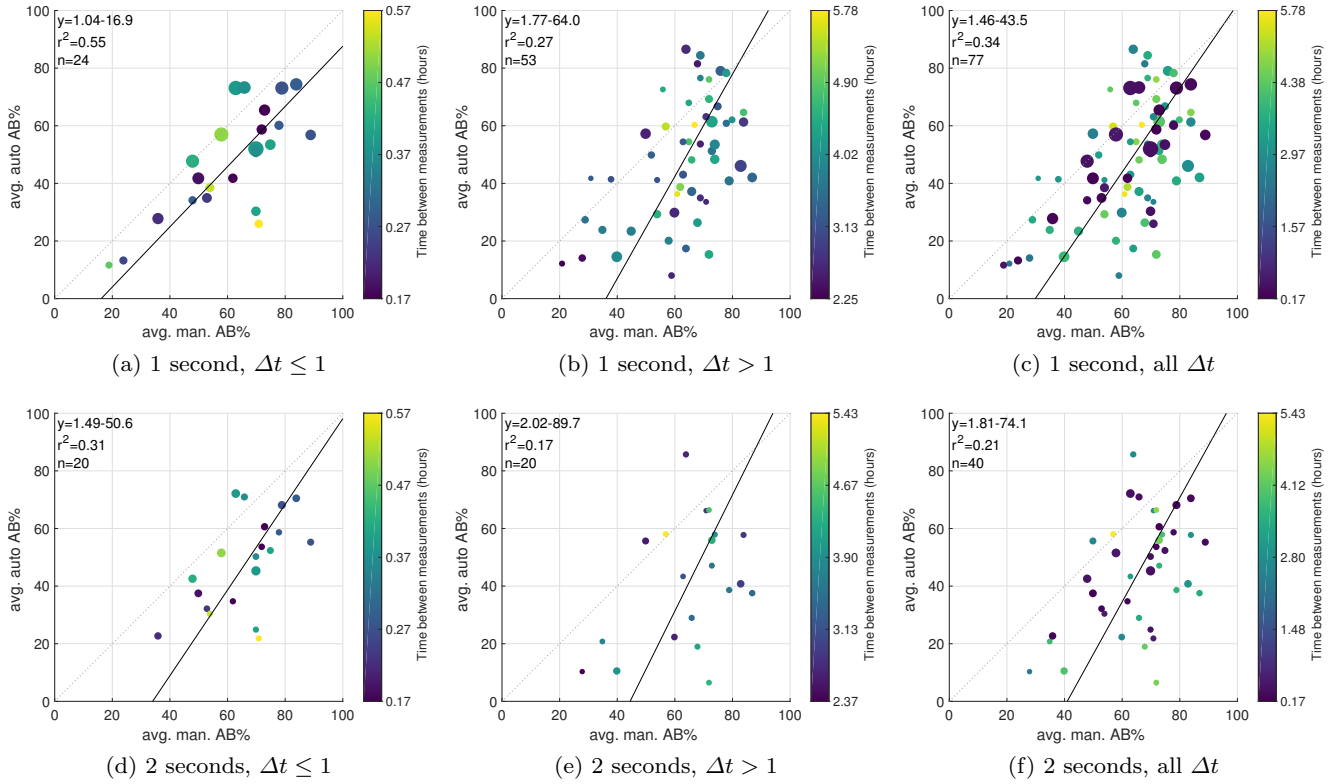
Table 1 AB/CD and C/D optimal threshold information for 25 and 50 frames analyses.

In summary, comparing automated with manual motility seems highly affected by how the automated measurements are translated into manual ABCD categories. With optimised thresholds we obtain very good linearly correlated graphs for AB% and D% between manual and automatic analysis but with an increasing amount of variation for samples with high Δt. The optimised AB and D thresholds reported are used throughout the remaining experiments.

3.2 Temporal variation

Automated motility assessment allows us to investigate the variation in read-outs as a function of the observation duration. We measure the stability of the read-outs for consecutive read-outs of the same observation duration and the relationship between read-outs of different observation durations.

Fig. 7 shows the temporal variation of AB% for each aliquot of seconds 1 and 2 (a) and seconds 1-2 and 3-4 (b). We observed a near linear relationship of both distributions with $r^2 = 0.99$ and $r^2 = 0.97$ and limits of agreement of 4.7 and 3.9 respectively. Table 2 shows an overview of bias and RPC_{np} for different combinations of 1- and 2-second intervals for which LIN and AB% was calculated. Notice there was a bias between 1 and 2 seconds for LIN with a p-value of 0.03-0.04 but not for AB%. RPC_{np} was very similar between 1- and 2-second combinations for LIN whereas it was slightly lower for 2-second combinations for AB%. There is likewise a larger RPC_{np} for 1- and 2-second combinations compared to only combining 1 or 2 seconds for both LIN and AB%. This implies that the choice between 1 and 2 seconds depends on the metric being measured. The LIN bias is significant between 1 and 2 second durations but the AB% bias is not significant. In general, very little difference was observed for AB% as a function of observation duration. AB% RPC_{np} is slightly smaller for 2 second than for 1 second durations.



(a) 1 second, $\Delta t \leq 1$ (b) 1 second, $\Delta t > 1$ (c) 1 second, all Δt

(d) 2 seconds, $\Delta t \leq 1$ (e) 2 seconds, $\Delta t > 1$ (f) 2 seconds, all Δt

Fig. 1 Scatter plot of AB% from avg. manual and avg. automatic analysis. Plots are shown for 1- and 2-second analysis and three splits of samples based on time between measurements Δt. The exact time between the automated and manual analysis is indicated by the colour scale. The size of the points indicates average number of tracks in the automatic analysis ranging from 104 to 1593.

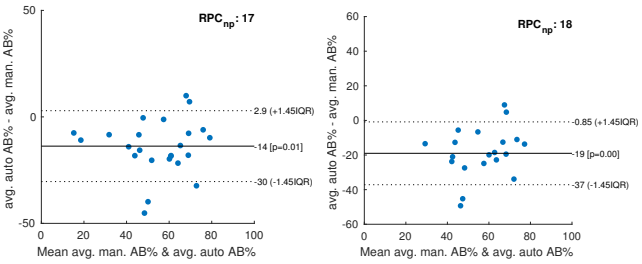
		LIN			AB%		
Time 1 (sec (frames))	Time 2 (sec (frames))	Bias	Bias p	RPC_{np}	Bias	Bias p	RPC_{np}
1 (1-25)	2 (26-50)	0.0002	1.00	0.0376	0.18	0.97	4.74
1 (1-25)	3 (51-75)	−0.0007	0.91	0.0354	0.15	0.96	5.46
1 (1-25)	4 (76-100)	−0.0031	0.86	0.0345	−0.16	0.93	4.41
2 (26-50)	3 (51-75)	0.0001	0.87	0.0324	−0.11	1.00	5.05
2 (26-50)	4 (76-100)	−0.0016	0.84	0.0374	−0.21	0.87	4.04
3 (51-75)	4 (76-100)	−0.0001	0.97	0.0334	−0.38	0.85	4.38
1 (1-25)	3-4 (51-100)	−0.0413	**0.03**	0.0472	2.10	0.27	8.01
2 (26-50)	3-4 (51-100)	−0.0351	**0.03**	0.0411	2.13	0.30	7.48
3 (51-75)	1-2 (1-50)	−0.0368	**0.04**	0.0424	2.50	0.30	7.72
4 (76-100)	1-2 (1-50)	−0.0362	**0.03**	0.0463	2.44	0.31	8.16
1-2 (1-50)	3-4 (51-100)	0.0013	0.99	0.0369	0.17	0.92	3.92

Table 2 Summary of Bland-Altman analysis for combinations of selected non-overlapping 1- and 2-second timeslots. Bias p-values of less than 0.05 are highlighted in bold indicating a 95% significance level.

3.3 Intra-aliquot variation

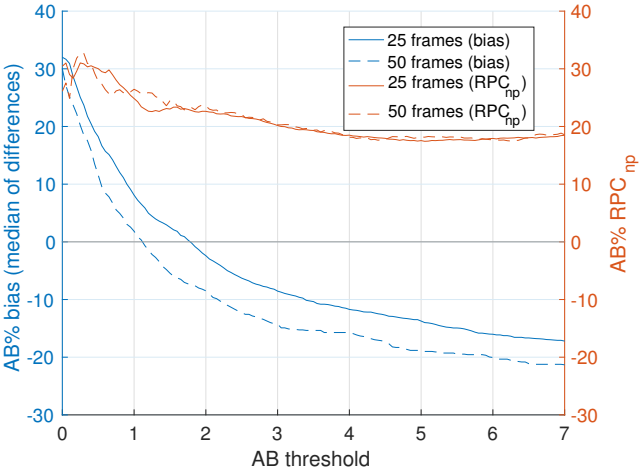
We subsequently investigate the variation in read-outs within each aliquot. The experiment gives an estimate of the confidence we have in the read-outs from the automated method for semen under the same physical conditions and handled exactly the same. Views from each aliquot were split into two groups with a close to equal amount of tracked sperm cells and the resulting motility estimations were compared.

Fig. 8 shows BA plots for the AB% variation across the view groups for 1 (Fig. 8a) and 2 (Fig. 8b) seconds. Only samples with at least 100 tracked sperm cells in each view split have been included for reliable estimates of AB%. The distribution follows a general linear tendency with few outliers. RPC_{np} is the same for 1 (14) and 2 (14) second videos and is higher than the tempo-

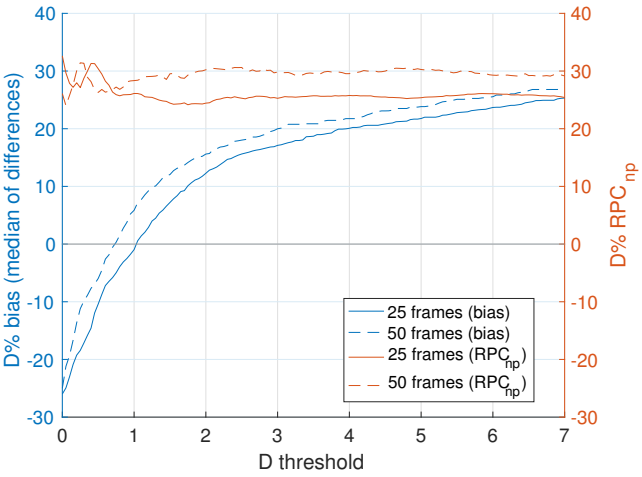


(a) 1 second, $\Delta t \leq 1$ (b) 2 seconds, $\Delta t \leq 1$

Fig. 2 Bland-Altman plot of AB% from avg. manual and avg. automatic analysis. Samples with at least 2x100 sperm cell tracks are included.



(a) AB threshold



(b) D threshold

Fig. 3 Plots of bias (blue) and $\mathrm{RPC_{np}}$(orange) measured on samples with $\Delta t \leq 1$ when varying the AB (a) and D (b) thresholds within 0–7 µm/s with steps of 0.05 µm/s. Each plot shows data for analysis of both 25 frames (line) and 50 frames (dashed line).

ral variation (4.7 and 3.9 for 1- and 2-second analysis, respectively).

3.4 Inter-aliquot variation

The manual WHO motility analysis procedure dictates that two aliquots should be scored and compared to increase the certainty of the analysis. We estimate the variation of read-outs from multiple aliquots of semen handled according to the same protocol for both automatic and manual analysis.

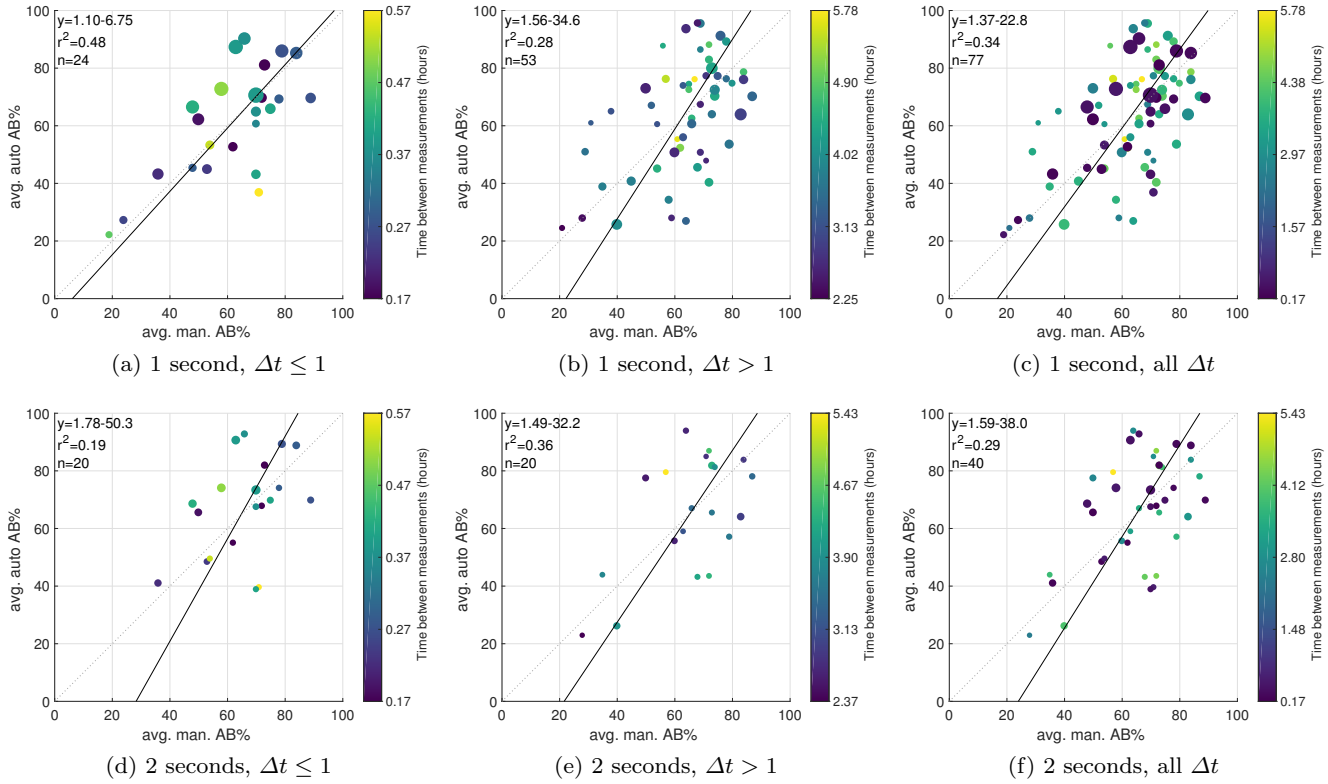
The BA plots for the inter-aliquot variation estimation of AB% are shown in Fig. 14 and 12 for the automatic and manual motility analysis respectively. The automatic results are shown for both 1 and 2 seconds as well as two different inclusion criteria: samples having a minimum of 100 sperm tracks in each aliquot (Fig. 14 a-b) and samples living up to the same quality control criteria as the manual assessments (Fig. 14 c-d). Similar plots are found in the supplementary material for motility grades C and D. The plots for automatic motility estimation (Fig. 14 a-b) show a linear correlation for 59 and 37 1- and 2-second analysis samples, respectively. For the remaining 18 and 3 1- and 2-second analysis, respectively, we see a considerably lower motility in aliquot 2 compared to aliquot 1. Samples conforming to similar quality control criteria as manually determined read-outs naturally show a very good linear correlation with r^2 values of 0.97 and $\mathrm{RPC_{np}}$of 10 and 7.0 for 1- and 2-second analysis respectively (Fig. 14 c-d). The manual analysis (Fig. 12) shows a linear correlation with $r^2 = 0.92$ and a $\mathrm{RPC_{np}}$ of 9.1. Notice the difference in the number of samples included after the quality control check between automatic (18 and 14 for 1- and 2-second analyses, respectively) and manual (77) motility analysis.

Take together, the observed variation within each sample was considerable and when forcing read-outs to obey the Poisson distribution automated evaluation performed slightly better than manual evaluation.

4 Discussion

4.1 Comparison with manual motility

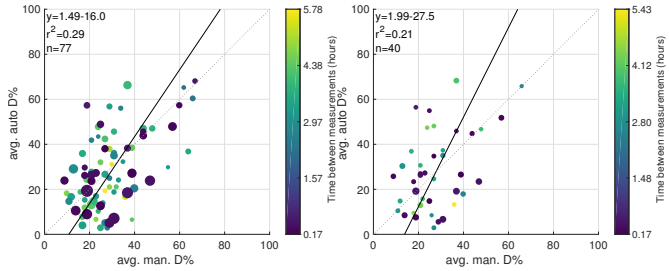
Due to the large effect of time on motility assessments we divided our data into two categories when comparing with the manual analysis: Correlation for samples with less than one hour between the analyses and with more than one hour between the analyses. Comparison of both these sub-parts of the dataset showed correlations as expected: Analyses that were conducted tem-



(a) 1 second, $\Delta t \leq 1$ (b) 1 second, $\Delta t > 1$ (c) 1 second, all Δt

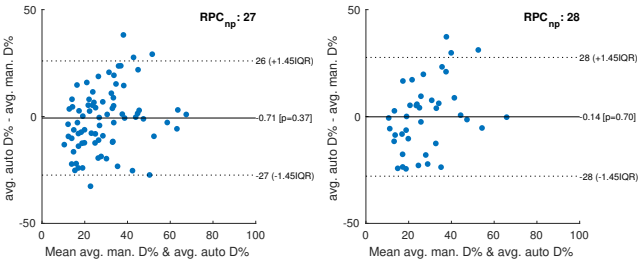
(d) 2 seconds, $\Delta t \leq 1$ (e) 2 seconds, $\Delta t > 1$ (f) 2 seconds, all Δt

Fig. 4 Scatter plot of AB% from avg. manual and avg. automatic analysis. Plots are shown for 1- and 2-second analysis and three splits of samples based on time between measurements Δt. The exact time between the automated and manual analysis is indicated by the colour scale. The size of the points indicates average number of tracks in the automatic analysis ranging from 104 to 1593.



(a) 1 second, all Δt (b) 2 seconds, all Δt

Fig. 5 Scatter plot of D% from avg. manual and avg. automatic analysis for 1- and 2-second analysis. The marker size indicates average number of tracks in the automatic analysis ranging from 104 to 1593.

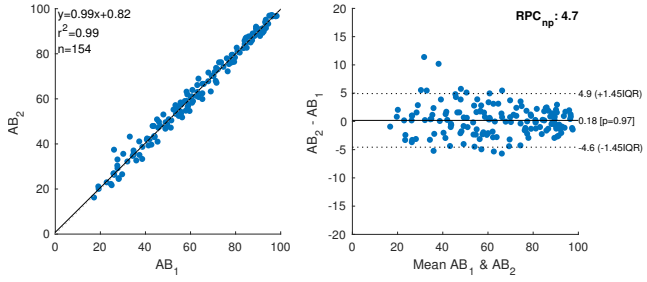


(a) 1 second, all Δt (b) 2 seconds, all Δt

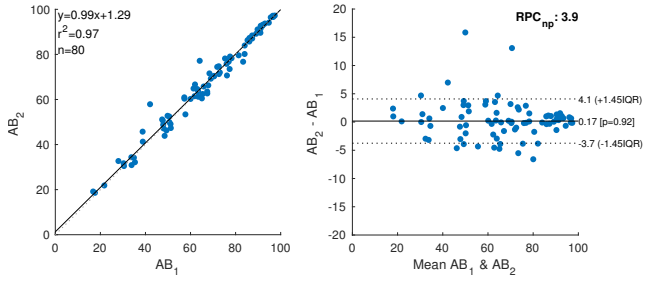
Fig. 6 Bland-Altman plot of D% from avg. manual and avg. automatic analysis with optimised D threshold. Samples with at least 2x100 sperm cell tracks are included.

porally close to each other had a good linear correlation with a few outliers. The analyses conducted temporally further apart were scattered more towards lower motilities for the last (automatic) analysis due to the temporal decrease of motility affecting samples with varying effect.

The temporally close analyses had a bias towards lower motility for the automatic analysis. We therefore adjusted the VSL gating parameters to minimise the bias between the two methods for samples with $\Delta t \leq 1$ making our automatic analysis reflect the manual analysis. The optimisation of bias indicated, that we needed different thresholds for 1- and 2-second analysis. This difference could be caused by differences in metric readout (VSL) for the single sperm cell caused by the difference in duration as Nissen et al. [12] observed for LIN. This argument is supported by the bias in AB% observed between 1- and 2-second analysis, though the bias was non-significant.

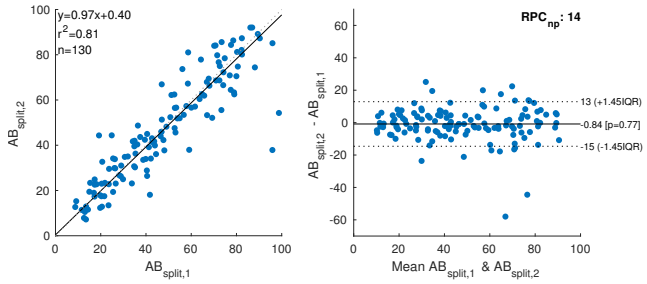


(a) Second 1 vs. 2

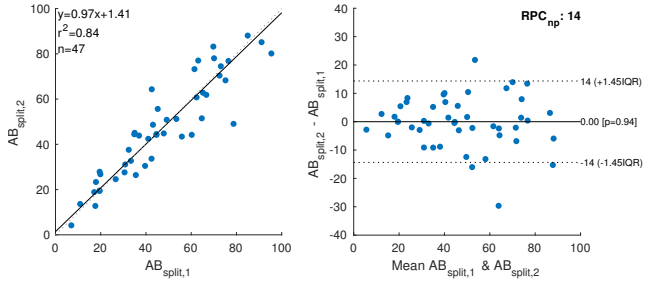


(b) Seconds 1-2 vs. 3-4

Fig. 7 Bland-Altman plots of AB% computed based on non-overlapping framesubsets for the same aliquots.



(a) 1 second (frames 1-25)



(b) 2 seconds (frames 1-50)

Fig. 8 Bland-Altman plot of AB% computed based on view splits 1 and 2 for automatic counting.

Having statistically comparable measurements between manual and automatic analysis, we could change to using the automatic analysis instead of the manual. Theoretically, there is one big advantage of choosing CASA over manual analysis: sperm cell sample size. When conducting motility estimation we estimate the global motility distribution in a semen sample from a small subset of observed sperm cells. Assume we have a method that accurately estimates the motility of a single sperm cell. The accuracy of the global motility estimate increases as our sperm cell sample size increases. With CASA we can increase the sperm cell sample size compared to the manual analysis without increasing the laboratory technician hands-on time, thereby also increasing the global motility estimate accuracy.

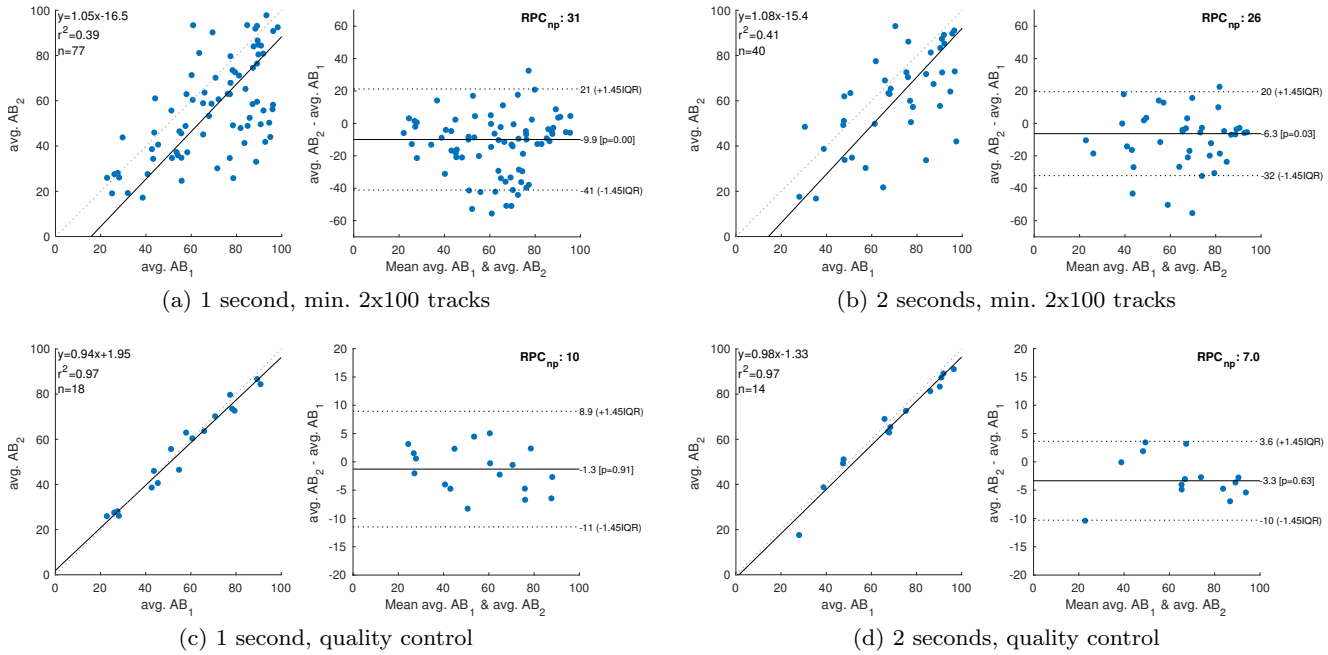
4.2 Temporal variation

Our data on temporal variation verified our previous findings of a LIN bias between 1- and 2-second analysis as reported in Nissen et al. [12], whereas there was no significant bias for AB% between 1- and 2-second analysis. The RPC_{np} for AB% was bigger for 1- than 2-second analysis, which was not the case for LIN. We need to be careful with the choice of metric and design of data analysis due to these differences in metric behaviour/dependencies.

Our data also showed the existence of temporal variation in the metrics we computed limiting the certainty of each automatic motility estimation. Briefly explained we are trying to estimate the global motility of a semen sample by sampling the motion of a subset of the sperm cells in the sample from a short duration. Even small abrupt behaviour expressed by a single sperm cell has the potential to alter the read-outs of very specific and precise metrics. This is however also the case with manual assessment of motility.

4.3 Intra-aliquot variation

The results of the intra-aliquot variation experiment indicated, that the variation across subsets of views in the same aliquot is larger than the temporal variation in the full set of views. Some of the increased variation most likely originates from estimating the motility distribution from half the number of sperm cell tracks decreasing the accuracy of the estimate. We only included aliquots/samples with view splits having at least 2x100 tracked sperm cells in order to maintain a decent level of accuracy of the estimated motility. Another part of the increased variation might come from inconsistent views due to a non-uniform spatial sperm cell distribution. Our view splits were made based on sperm track count in each view, splitting up the views such that we achieved a minimal difference between the total number of cells in each subgroup. In case of inconsistent motility of views this could lead to inconsistency of motility in the subgroups. For example: One subgroup could consist of one view with a large number of sperm cells with



(a) 1 second, min. 2x100 tracks

(b) 2 seconds, min. 2x100 tracks

(c) 1 second, quality control

(d) 2 seconds, quality control

Fig. 9 Bland-Altman plots for the two aliquots AB_1 and AB_2 for automatic counting for 1 and 2 seconds. Two different sample inclusion criteria are used: samples with minimum 2x100 sperm cell tracks and samples with aliquots living up to the quality control criteria.

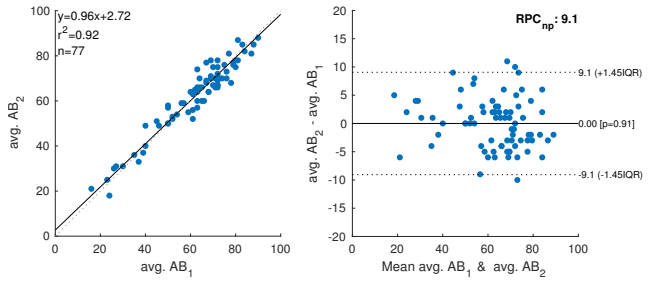


Fig. 10 Bland-Altman plot of the two aliquots AB_1 and AB_2 for manual counting of 2x100 sperm cells.

high motility, and the other subgroup could consist of a large number of views each with fewer and less motile sperm cells.

4.4 Inter-aliquot variation

The inter-aliquot variation experiments with relaxed inclusion criteria (2x100 tracks) clearly indicated inconsistencies between a subgroup of the samples while the remaining samples showed a good inter-aliquot correlation. The subgroup of inconsistent samples all had a higher motility in aliquot 1 than 2 (except one). Aliquot 1 was always the first aliquot being captured, and the subgroup therefore indicated a systematic problem with our setup affecting the motility analysis of aliquot 2 for certain samples. These problems could include: sample drift, temperature changes, pipette handling, dilution technique, natural motility decrease, and mixing of samples.

Adding the quality control criteria made us filter away a very high number of the samples (59/77 and 26/40 for 1- and 2-second analyses respectively), for which new analysis would have to be conducted if we were to use the same criteria as the manual estimation. The remaining samples were nicely correlated as expected. The automatic analysis is able to analyse a higher number of sperm cells than the manual analysis. This gives us a more strict quality control criteria due to the definition of the Poisson distribution. We could ask, if the model for quality control is reasonable given the amount of both temporal and intra-aliquot variation? Do we risk throwing away valuable information in the process of trying to fit the data to our model?

The manual analysis with quality control had a slightly higher $\mathrm{RPC_{np}}$ compared to the automatic analysis. This result was expected, as the quality control automatically excludes all results that would lead to higher variance between aliquots and that the automatic method has better counting statistics.

The automatic system requires no interaction during analysis freeing up the time spent on counting by the technician. This will allow the technician to focus on sample handling minimising the intrinsic vari-

ation due to sample drift, pipette handling, and non-representative aliquots.

One of the reasons for differences between aliquots could be caused by sample handling. It can be difficult to get two consistent aliquots from a semen sample if the semen is inconsistent (lumpy) or if it has high viscosity. Drift can also be very difficult to avoid in combination with avoiding air bubbles (these will create local drift). In order to avoid air bubbles one can let the cover slip touch the droplet at an acute angle from the side before dropping the rest of the cover slip onto the droplet. This technique has a tendency of creating a global drift due to the angulation of the cover slip. The issue with drift is, that immotile cells move thus making it difficult to get a good estimate of the actual motility distribution. Unfortunately the current automatic analysis does not handle any level of drift whereas manual analysis is more robust against a light amount of drift. The issue with handling drift automatically is that drift can be very local in non-consistent samples and that immotile cells may stick to the glass slide or cover slip. Therefore we cannot estimate a global drift and transform all tracked paths accordingly. A different option recommended by WHO [15, pp. 21, Section 2.5] is to wait for the sample to stop drifting. We followed this recommendation in case of light drift, but if a heavy drift was observed, the cells would typically drift out from underneath the cover slip before the drift settled. This left very few sperm cells to be observed, and we generally observed a shift towards lower motility in the remaining cells due to an overrepresentation of immotile cells sticking to the slide or cover slip.

4.5 Duration of analysis

We conducted all our experiments for both 1- and 2-second analysis. In general, the 2-second analysis excludes more tracks than the 1-second analysis in order to obtain an unbiased motility estimate as described by Nissen et al. [**?**]. This caused fewer sample to live up to the requirement of having a least 2x100 sperm in our experiments after analysis. In order to avoid this issue, it would be advisable to capture more views of the samples ensuring a sufficient amount of sperm cell tracks to base the analysis on. We observed a smaller temporal variation for 2- than 1-second analysis indicating that we obtain better/more consistent statistics for the population when observing it for 2 seconds rather than 1 second. In conclusion, 2-second analysis achieved the most consistent results with the added cost of having to capture more views than for 1-second analysis.

5 Conflicts of Interest

The manufacturer of the image cytometer used in our study was a partner in the jointly funded project. All experiments and evaluations of results were conducted without influence from the funding bodies.

References

1. Belsey, M., Moghissi, K., Eliasson, R., Paulsen, C., Gallegos, A., Prasad, M.: Laboratory manual for the examination of human semen and semen-cervical mucus interaction. (1980)
2. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet **327**(8476), 307 – 310 (1986). DOI https://doi.org/10.1016/S0140-6736(86)90837-8. URL `http://www.sciencedirect.com/science/article/pii/S0140673686908378`. Originally published as Volume 1, Issue 8476
3. Cooper, T.G., Yeung, C.H.: Computer-aided evaluation of assessment of "grade a" spermatozoa by experienced technicians. Fertility and Sterility **85**(1), 220–224 (2006). DOI https://doi.org/10.1016/j.fertnstert.2005.07.1286. URL `http://www.sciencedirect.com/science/article/pii/S0015028205034308`
4. Egeberg, D., Kjærulff, S., Hansen, C., Petersen, J., Glensbjerg, M., Skakkebæk, N., Jørgensen, N., Almstrup, K.: Image cytometer method for automated assessment of human spermatozoa concentration. Andrology **1**(4), 615–623 (2013)
5. Egeberg Palme, D.L., Rehfeld, A., Bang, A.K., Nikolova, K.A., Kjærulff, S., Petersen, M.R., Jeppesen, J.V., Glensbjerg, M., Juul, A., Skakkebæk, N.E., Ziebe, S., Jørgensen, N., Almstrup, K.: Viable acrosome-intact human spermatozoa in the ejaculate as a marker of semen quality and fertility status. Human Reproduction **33**(3), 361–371 (2018). DOI 10.1093/humrep/dex380. URL `http://dx.doi.org/10.1093/humrep/dex380`
6. Goodson, S.G., White, S., Stevans, A.M., Bhat, S., Kao, C.Y., Jaworski, S., Marlowe, T.R., Kohlmeier, M., McMillan, L., Zeisel, S.H., O'Brien, D.A.: Casanova: a multiclass support vector machine model for the classification of human sperm motility patterns†. Biology of Reproduction **97**(5), 698–708 (2017). DOI 10.1093/biolre/iox120. URL `http://dx.doi.org/10.1093/biolre/iox120`
7. Hidayatullah, P., Awaludin, I., Kusumo, R.D., Nuriyadi, M.: Automatic sperm motility measurement. In: 2015 International Conference on Information Technology Systems and Innovation (ICITSI), pp. 1–5 (2015). DOI 10.1109/ICITSI.2015.7437674
8. van der Horst, G., Mortimer, S.T., Mortimer, D.: The future of computer-aided sperm analysis. Asian Journal of Andrology **17**(4), 545–553 (2015). DOI 10.4103/1008-682X.154312
9. Mortimer, D., Mortimer, S.T.: Laboratory investigation of the infertile male. A Textbook of In-Vitro Fertilization and Assisted Reproduction. 3rd ed. London: Taylor and Francis Medical Books pp. 61–91 (2005)
10. Mortimer, D., Serres, C., Mortimer, S.T., Jouannet, P.: Influence of image sampling frequency on the perceived movement characteristics of progressively motile human spermatozoa. Gamete Research **20**(3), 313–327 (1988).

DOI 10.1002/mrd.1120200307. URL http://dx.doi.org/10.1002/mrd.1120200307
11. Nissen, M.S., Krause, O., Almstrup, K., Kjærulff, S., Nielsen, T.T., Nielsen, M.: Convolutional Neural Networks for Segmentation and Object Detection of Human Semen, pp. 397–406. Springer International Publishing, Cham (2017). DOI 10.1007/978-3-319-59126-1_33. URL http://dx.doi.org/10.1007/978-3-319-59126-1_33
12. Nissen, M.S., Krause, O., Almstrup, K., Kjærulff, S., Meijering, E., Nielsen, M.: Estimation of motility distribution: A study of linearity for human sperm motility analysis (2018). Journal paper to be submitted for Transactions on Medical Imaging (T-MI)
13. Palme, D.L.E., Johannsen, T.H., Petersen, J.H., Skakkebæk, N.E., Juul, A., Jørgensen, N., Almstrup, K.: Validation of image cytometry for sperm concentration measurement: Comparison with manual counting of 4010 human semen samples. Clinica Chimica Acta **468**(Supplement C), 114 – 119 (2017). DOI https://doi.org/10.1016/j.cca.2017.02.014. URL http://www.sciencedirect.com/science/article/pii/S0009898117300669
14. World Health Organisation: WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction. Cambridge university press (1999)
15. World Health Organization, et al.: WHO laboratory manual for the examination and processing of human semen (2010)
16. York, D.: Least-squares fitting of a straight line. Canadian Journal of Physics **44**(5), 1079–1086 (1966)

A Supplementary material

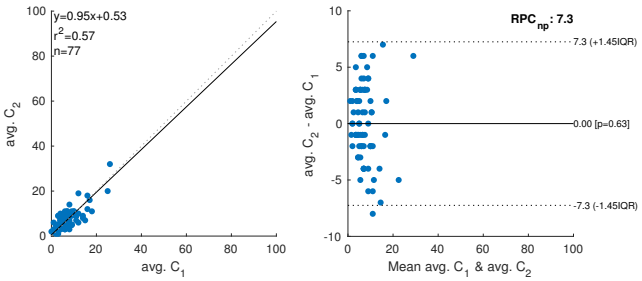


Fig. 11 Bland-Altman plot of the two aliquots C_1 and C_2 for manual counting of 2x100 sperm cells.

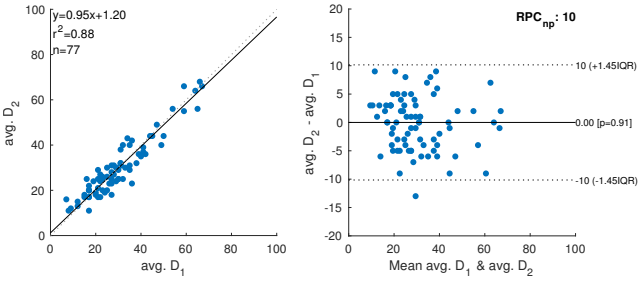
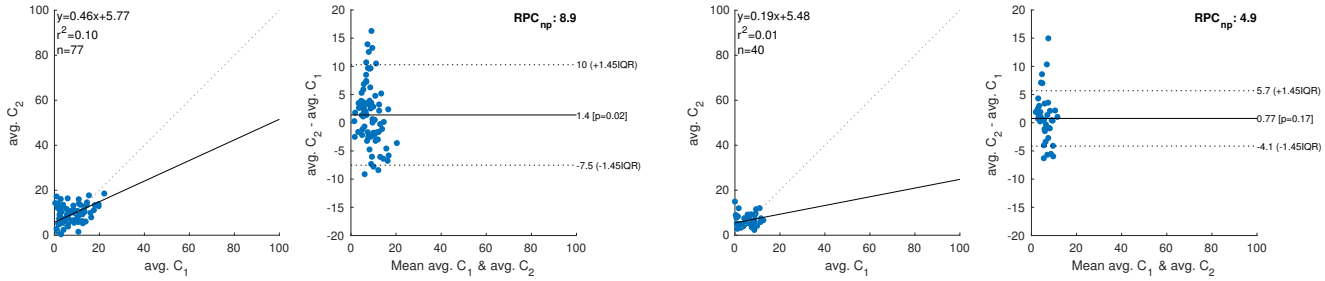
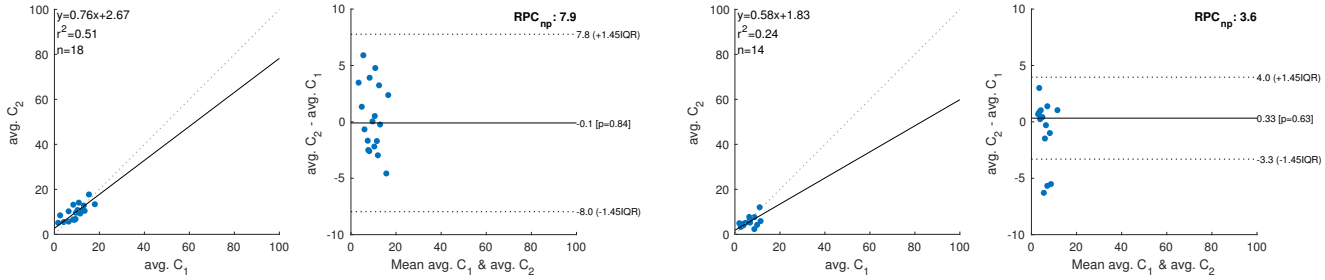


Fig. 12 Bland-Altman plot of the two aliquots D_1 and D_2 for manual counting of 2x100 sperm cells.

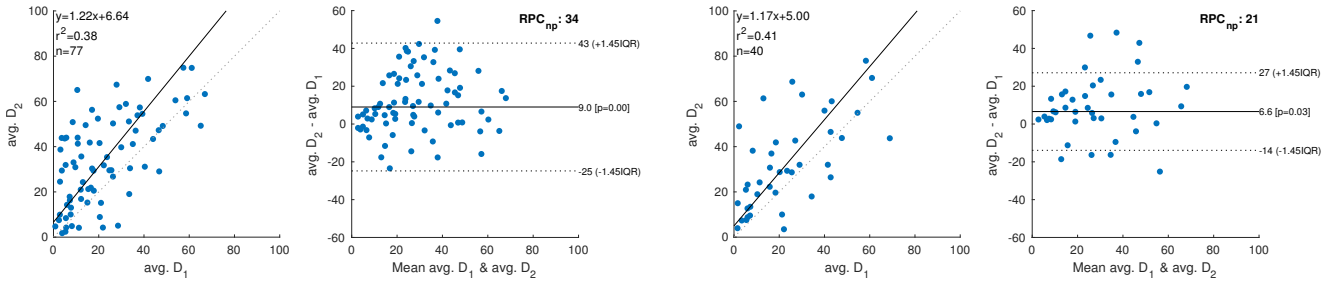


(a) 1 second, min. 2x100 tracks (b) 2 seconds, min. 2x100 tracks

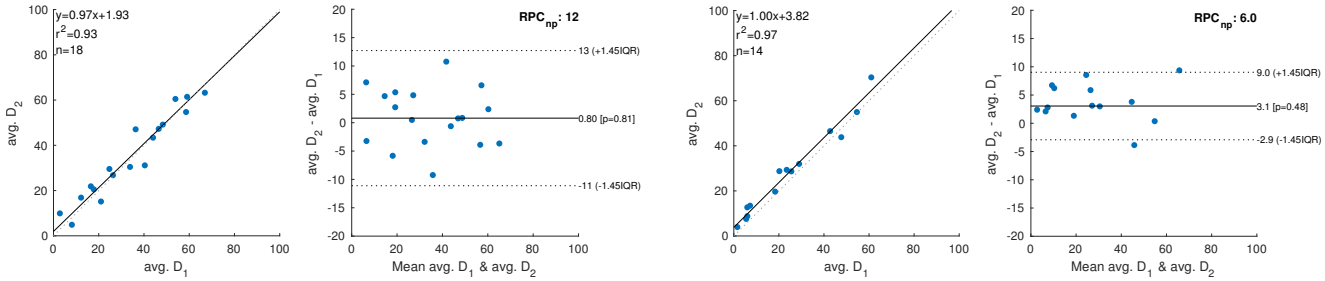


(c) 1 second, quality control (d) 2 seconds, quality control

Fig. 13 Bland-Altman plots for the two aliquots C_1 and C_2 for automatic counting for 1 and 2 seconds. Two different sample inclusion criteria are used: samples with minimum 2x100 sperm cell tracks and samples with aliquots living up to the quality control criteria.



(a) 1 second, min. 2x100 tracks (b) 2 seconds, min. 2x100 tracks



(c) 1 second, quality control (d) 2 seconds, quality control

Fig. 14 Bland-Altman plots for the two aliquots D_1 and D_2 for automatic counting for 1 and 2 seconds. Two different sample inclusion criteria are used: samples with minimum 2x100 sperm cell tracks and samples with aliquots living up to the quality control criteria.

Bibliography

[1] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*. Vol. 4. AMLBook New York, NY, USA: 2012.

[2] Lu J. C., Huang Y. F., and Lü N. Q. "Computer-aided sperm analysis: past, present and future". In: *Andrologia* 46.4 (Apr. 2013), pp. 329–338. DOI: 10.1111/and.12093. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/and.12093. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/and.12093.

[3] D. Ciresan, U. Meier, and J. Schmidhuber. "Multi-column deep neural networks for image classification". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. June 2012, pp. 3642–3649. DOI: 10.1109/CVPR.2012.6248110.

[4] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. "Mitosis detection in breast cancer histology images with deep neural networks". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer, 2013, pp. 411–418.

[5] Trevor G. Cooper and Ching-Hei Yeung. "Computer-aided evaluation of assessment of "grade a" spermatozoa by experienced technicians". In: *Fertility and Sterility* 85.1 (2006), pp. 220–224. ISSN: 0015-0282. DOI: https://doi.org/10.1016/j.fertnstert.2005.07.1286. URL: http://www.sciencedirect.com/science/article/pii/S0015028205034308.

[6] Jifeng Dai, Kaiming He, and Jian Sun. "Instance-aware Semantic Segmentation via Multi-task Network Cascades". In: *CoRR* abs/1512.04412 (2015). URL: http://arxiv.org/abs/1512.04412.

[7] Russel O. Davis and David F. Katz. "Standardization and Comparability of CASA Instruments". In: *Journal of Andrology* 13.1 (1992), pp. 81–86. ISSN: 1939-4640. DOI: 10.1002/j.1939-4640.1992.tb01632.x. URL: http://dx.doi.org/10.1002/j.1939-4640.1992.tb01632.x.

[8] Dorte Louise Egeberg Palme et al. "Viable acrosome-intact human spermatozoa in the ejaculate as a marker of semen quality and fertility status". In: *Human Reproduction* 33.3 (2018), pp. 361–371. DOI: 10.1093/humrep/dex380. eprint: /oup/backfile/content_public/journal/humrep/33/3/10.1093_humrep_dex380/1/dex380.pdf. URL: http://dx.doi.org/10.1093/humrep/dex380.

[9] DL Egeberg et al. "Image cytometer method for automated assessment of human spermatozoa concentration". In: *Andrology* 1.4 (2013), pp. 615–623.

[10] Fatemeh Ghasemian, Seyed Abolghasem Mirroshandel, Sara Monji-Azad, Mahnaz Azarnia, and Ziba Zahiri. "An efficient method for automatic morphological abnormality detection from human sperm images". In: *Computer Methods and Programs in Biomedicine* 122.3 (2015), pp. 409–420. ISSN: 0169-2607. DOI: http://dx.doi.org/10.1016/j.cmpb.2015.08.013. URL: http://www.sciencedirect.com/science/article/pii/S0169260715002230.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[12] Summer G Goodson et al. "CASAnova: a multiclass support vector machine model for the classification of human sperm motility patterns†". In: *Biology of Reproduction* 97.5 (2017), pp. 698–708. DOI: 10.1093/biolre/iox120. eprint: /oup/backfile/content_public/journal/biolreprod/97/5/10.1093_biolre_iox120/3/iox120.pdf. URL: http://dx.doi.org/10.1093/biolre/iox120.

[13] Toft Gunnar, Rignell-Hydbom Anna, Tyrkiel Ewa, Shvets Maryna, and Giwercman Aleksander. "Quality control workshops in standardization of sperm concentration and motility assessment in multicentre studies". In: *International Journal of Andrology* 28.3 (2005), pp. 144–149. DOI: 10.1111/j.1365-2605.2005.00518.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2605.2005.00518.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2605.2005.00518.x.

[14] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit". In: *Nature* 405.6789 (2000), p. 947.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[16] Yuki Hirano et al. “ANDROLOGY: Relationships Between Sperm Motility Characteristics Assessed by the Computer-Aided Sperm Analysis (CASA) and Fertilization Rates In Vitro”. In: *Journal of Assisted Reproduction and Genetics* 18.4 (Apr. 2001), pp. 215–220. ISSN: 1573-7330. DOI: 10.1023/A:1009420432234. URL: https://doi.org/10.1023/A:1009420432234.

[17] Gerhard van der Horst, Sharon T Mortimer, and David Mortimer. “The future of computer-aided sperm analysis”. In: *Asian Journal of Andrology* 17.4 (2015), pp. 545–553. DOI: 10.4103/1008-682X.154312.

[18] A. Jaiswal, W. J. Godinez, R. Eils, M. J. Lehmann, and K. Rohr. “Tracking Virus Particles in Fluorescence Microscopy Images Using Multi-Scale Detection and Multi-Frame Association”. In: *IEEE Transactions on Image Processing* 24.11 (Nov. 2015), pp. 4122–4136. ISSN: 1057-7149. DOI: 10.1109/TIP.2015.2458174.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[20] Lene Larsen et al. “Computer-assisted semen analysis parameters as predictors for fertility of men from the general population”. In: *Human Reproduction* 15.7 (2000), pp. 1562–1567. DOI: 10.1093/humrep/15.7.1562. eprint: /oup/backfile/content_public/journal/humrep/15/7/10.1093_humrep_15.7.1562/1/0151562.pdf. URL: http://dx.doi.org/10.1093/humrep/15.7.1562.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), p. 436.

[22] Yann LeCun et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551.

[23] Hagai Levine et al. “Temporal trends in sperm count: a systematic review and meta-regression analysis”. In: *Human Reproduction Update* 23.6 (2017), pp. 646–659. DOI: 10.1093/humupd/dmx022. eprint: /oup/backfile/content_public/journal/humupd/23/6/10.1093_humupd_dmx022/1/dmx022.pdf. URL: http://dx.doi.org/10.1093/humupd/dmx022.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

[25] Serdia O. Mack, Don P. Wolf, and Joseph S. Tash. "Quantitation of Specific Parameters of Motility in Large Numbers of Human Sperm by Digital Image Processing". In: *Biology of Reproduction* 38.2 (1988), pp. 270–281. DOI: 10.1095/biolreprod38.2.270. eprint: /oup/backfile/content_public/journal/biolreprod/38/2/10.1095_biolreprod38.2.270/1/biolreprod0270.pdf. URL: +%20http://dx.doi.org/10.1095/biolreprod38.2.270.

[26] Mary Condon Mahony, Nancy J Alexander, and R James Swanson. "Evaluation of semen parameters by means of automated sperm motion analyzers". In: *Fertility and sterility* 49.5 (1988), pp. 876–880.

[27] Maya N. Mascarenhas, Seth R. Flaxman, Ties Boerma, Sheryl Vanderpoel, and Gretchen A. Stevens. "National, Regional, and Global Trends in Infertility Prevalence Since 1990: A Systematic Analysis of 277 Health Surveys". In: *PLOS Medicine* 9.12 (Dec. 2012), pp. 1–12. DOI: 10.1371/journal.pmed.1001356. URL: https://doi.org/10.1371/journal.pmed.1001356.

[28] D. Mortimer, C. Serres, S. T. Mortimer, and P. Jouannet. "Influence of image sampling frequency on the perceived movement characteristics of progressively motile human spermatozoa". In: *Gamete Research* 20.3 (1988), pp. 313–327. ISSN: 1554-3919. DOI: 10.1002/mrd.1120200307. URL: http://dx.doi.org/10.1002/mrd.1120200307.

[29] David Mortimer. "Laboratory standards in routine clinical andrology". In: *Reproductive Medicine Review* 3.2 (1994), pp. 97–111. DOI: 10.1017/S0962279900000818.

[30] Malte S. Nissen, Oswin Krause, Kristian Almstrup, Søren Kjærulff, Erik Meijering, and Mads Nielsen. "Estimation of motility distribution: A study of linearity for human sperm motility analysis". Journal paper to be submitted for Transactions on Medical Imaging (T-MI). 2018.

[31] Malte S. Nissen, Oswin Krause, Kristian Almstrup, Søren Kjærulff, Torben T. Nielsen, and Mads Nielsen. "Convolutional Neural Networks for Segmentation and Object Detection of Human Semen". In: *Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12–14, 2017, Proceedings, Part I*. Ed. by Puneet Sharma and Filippo Maria Bianchi. Cham: Springer International Publishing, 2017, pp. 397–406. ISBN: 978-3-319-59126-1. DOI: 10.1007/978-3-319-59126-1_33. URL: http://dx.doi.org/10.1007/978-3-319-59126-1_33.

[32] Malte S. Nissen, Mads Nielsen, and Kristian Almstrup. "Evaluation of a new integrated and fully automated system for sperm motility analysis". Journal paper to be submitted for Andrology. 2018.

[33] Jouannet P., Docut B., Feneux D, and Spira A. "Male factors and the likelihood of pregnancy in infertile couples. I. Study of sperm characteristics". In: *International Journal of Andrology* 11.5 (), pp. 379–394. DOI: 10.1111/j.1365-2605.1988.tb01011.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2605.1988.tb01011.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2605.1988.tb01011.x.

[34] Dorte L Egeberg Palme et al. "Validation of image cytometry for sperm concentration measurement: Comparison with manual counting of 4010 human semen samples". In: *Clinica Chimica Acta* 468 (2017), pp. 114–119.

[35] Pedro H. O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. "Learning to Refine Object Segments". In: *CoRR* abs/1603.08695 (2016). URL: http://arxiv.org/abs/1603.08695.

[36] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollar. "Learning to Segment Object Candidates". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 1990–1998. URL: http://papers.nips.cc/paper/5852-learning-to-segment-object-candidates.pdf.

[37] A. Rehfeld, S. Dissing, and N. E. Skakkebæk. "Chemical UV Filters Mimic the Effect of Progesterone on Ca2+ Signaling in Human Sperm Cells". In: *Endocrinology* 157.11 (2016), pp. 4297–4308. DOI: 10.1210/en.2016-1473. eprint: /oup/backfile/content_public/journal/endo/157/11/10.1210_en.2016-1473/6/endo4297.pdf. URL: +%20http://dx.doi.org/10.1210/en.2016-1473.

[38] Anders Rehfeld, Dorte Egeberg, Kristian Almstrup, Jørgen Holm Petersen, Steen Dissing, and Niels Erik Skakkebæk. "Chemical UV filters can affect human sperm function in a progesterone-like manner". In: *Endocrine Connections* (2017). DOI: 10.1530/EC-17-0156. eprint: http://www.endocrineconnections.com/content/early/2017/09/05/EC-17-0156.full.pdf+html. URL: http://www.endocrineconnections.com/content/early/2017/09/05/EC-17-0156.abstract.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 234–241.

[40] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), p. 533.

[41] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[42] Christian Schiffer et al. "Direct action of endocrine disrupting chemicals on human sperm". In: *EMBO reports* 15.7 (2014), pp. 758–765. ISSN: 1469-221X. DOI: 10.15252/embr.201438869. eprint: http://embor.embopress.org/content/15/7/758.full.pdf. URL: http://embor.embopress.org/content/15/7/758.

[43] K. Shafique and M. Shah. "A noniterative greedy algorithm for multiframe point correspondence". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.1 (Jan. 2005), pp. 51–65. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2005.1.

[44] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[45] Niels E. Skakkebaek et al. "Male Reproductive Disorders and Fertility Trends: Influences of Environment and Genetic Susceptibility". In: *Physiological Reviews* 96.1 (2016). PMID: 26582516, pp. 55–97. DOI: 10.1152/physrev.00017.2015. eprint: https://doi.org/10.1152/physrev.00017.2015. URL: https://doi.org/10.1152/physrev.00017.2015.

[46] Ihor Smal and Erik Meijering. "Quantitative comparison of multiframe data association techniques for particle tracking in time-lapse fluorescence microscopy". In: *Medical Image Analysis* 24.1 (2015), pp. 163–189. ISSN: 1361-8415. DOI: http://dx.doi.org/10.1016/j.media.2015.06.006. URL: http://www.sciencedirect.com/science/article/pii/S1361841515000936.

[47] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition". In: *Neural Networks* 32 (2012). Selected Papers from {IJCNN} 2011, pp. 323–332. ISSN: 0893-6080. DOI: http://dx.doi.org/10.1016/j.neunet.2012.02.016. URL: http://www.sciencedirect.com/science/article/pii/S0893608012000457.

[48] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[49] L. F. Urbano, P. Masson, M. VerMilyea, and M. Kam. "Automatic Tracking and Motility Analysis of Human Sperm in Time-Lapse Images". In: *IEEE Transactions on Medical Imaging* 36.3 (Mar. 2017), pp. 792–801. ISSN: 0278-0062. DOI: 10.1109/TMI.2016.2630720.

[50] World Health Organisation. *WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction*. Cambridge university press, 1999.

[51] World Health Organization, et al. "WHO laboratory manual for the examination and processing of human semen". In: (2010).

[52] Sergey Zagoruyko et al. “A MultiPath Network for Object Detection”. In: *CoRR* abs/1604.02135 (2016). URL: http://arxiv.org/abs/1604.02135.

[53] YT Zhou and R Chellappa. “Computation of optical flow using a neural network”. In: *IEEE International Conference on Neural Networks*. Vol. 1998. 1988, pp. 71–78.