Using Deep Learning to Predict Motility and Morphology of Human Sperm

Steven Hicks[1, 2], Trine B. Haugen[2], Pål Halvorsen[1, 2] Michael Riegler[1, 3]
[1]SimulaMet, Norway [2]Oslo Metropolitan University, Norway
[3]Kristiania University College, Norway

ABSTRACT

In the Medico Task 2019, the main focus is to predict sperm quality based on videos and other related data. In this paper, we present the approach of team *LesCats* which is based on deep convolution neural networks, where we experiment with different data preprocessing methods to predict the morphology and motility of human sperm. The achieved results show that deep learning is a promising method for human sperm analysis. Out best method achieves a mean absolute error of 8.962 for the motility task and a mean absolute error of 5.303 for the morphology task.

1 INTRODUCTION

In an effort to explore how medical multimedia can be used to create high performing and efficient prediction algorithms, the Multimedia for Medicine (Medico) Task presents different use-cases [6] which challenge computer science researchers to explore a field which has much potential for real-world impact. This year's task differs from previous years as it focuses on the analysis of microscopic videos of human semen to assess the quality of sperm. The videos are taken from the open-source VISEM dataset [4]. The challenge presents three different tasks, of which we decided to focus on the tasks which are required in order to participate this years challenge, i.e., the *prediction of motility task* and the *prediction of morphology task*. The tasks themselves are further described in the overview paper [5].

2 APPROACH

Our approach is based on deep learning using deep convolutional neural networks (CNNs) to predict sperm motility and sperm morphology. All experiments aim to utilize the information in the videos to their fullest, yet still keeping the computational complexity low. The experiments can primarily be split into four distinct groups. Firstly (i), we combine multiple frames channel-wise using different stride values (distance between selected frames) and feed this directly into the deep neural network. Secondly (ii), we vary the number of frames used in each sample to see how this may effect the algorithms prediction performance. Thirdly (iii), we threshold the colors of each frame in an attempt to separate the spermatozoa bright color from the darker background, and use this information for prediction. Lastly (iv), we add the patient data to the video analysis to see how this may help in the prediction. Because morphology focuses more on the visual appearance of sperm than the movement, we opted to perform the threshold experiments only on the motility experiments. Internally, we experimented with a wide variety of configurations, but only submitted the best results as the official runs. In the following few sections, we will give a brief explanation of our experimental setup (common training configuration between each model and data preparation), and a more detailed description of each approach.

2.1 Experimental Setup

For each experiment, we use the Inception V3 [7] architecture for our deep learning model, which were trained for as long as it improved on the validation loss. This means that the models trained indefinitely until the mean absolute error did not improve over the last 100 epochs. Each model was trained with batch size of 16 using Nadam [3] to optimize the weights with a learning rate of 0.001. The models were implemented using the Keras [2] deep learning library with a TensorFlow [1] back-end. Each experiment was performed on what would be considered "consumer-grade" hardware, specifically, a desktop computer running Arch Linux with an Intel core i7 processor, 16 gigabytes of RAM, and an Nvidia GTX 1080Ti graphics card. As the videos in the provided dataset vary in length (ranging from 2 to 7 minutes), we extracted a number of clips (one clip is contains a sequence of frames) from each video before training. The clips were extracted from evenly spaced out intervals throughout the entire video, meaning we get a set of clips which accurately represent any given semen recording. For both the prediction of motility and the prediction of morphology task, we use ZeroR as a baseline to measure our results.

2.2 Frame Stride Experiments

For the methods which used different stride lengths to perform prediction on sperm quality, we performed a total of 10 different experiments. Stride in this context refers to the distance between two extracted frames within a clip. For example, using a stride length of 5 would select every fifth frame within a given frame sequence. The purpose of this experiment is to exaggerate the change between two frames by increasing the distance of where the two frames were sampled. Each experiment used a clip length of three frames which are greyscaled, resized to 224×224 pixels and combined channel-wise. The result is that each clip has a shape of 224×3, making it possible to use pre-trained networks. We take advantage of this attribute and train two models for each stride value tested, i.e., one transferring the weights of an ImageNet-based model and one trained from scratch. As previously stated, we performed a total of ten different experiments, of which five different stride values were used; 1, 5, 10, 30, and 50.

2.3 Clip Length Experiments

For the methods which used different clip lengths to predict sperm quality, we performed a total of 5 experiments. Each experiment

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

increases the number of frames in a clip by 10, starting at 10 and ending at 50. Each video is captured at 50 frames-per-second, which means that the clips which contain 50 frames represent a whole second of a given video. In contrast to the previous method, each clip included in these experiments have a stride of 1, meaning each frame in a sequence is used for prediction. Similar to the previous method, each frame resized to 224×224 and greyscaled before being combined channel-wise. The shape of each clip is then $224 \times 224 \times C$, where C is the length of the clip.

2.4 Threshold Experiments

For the threshold approach, we greyscale each extracted frame and threshold the color at 220, meaning all color values below 220 are set to 0. The spermatozoa in the provided videos have a strong bright coloring which differentiates it from the darker background. By thresholding the color values, we aim to separate the spermatozoa from the background in order to better emphasize the movement across frames. However, by doing this, we lose some of the visual information present in each sperm, that is why we chose not to apply this method to predict morphology. We organize these experiments in a similar manner as those done for the stride experiments, meaning we stack three frames channel-wise using five different stride values; 1, 5, 10, 25, and 50.

3 RESULTS AND DISCUSSION

Each method was evaluated using three-fold cross-validation (as required by the task), and we report the mean absolute error (MAE) and mean absolute error (RMSE) for each experiment. The results for the motility experiments are shown in Table 1, and the results for the morphology experiments are shown in Table 2.

As we can see the prediction of motility results (Table 1), using larger strides between the selected frames in combination with transfer learning works best. The experiments which used a lot of frames per clip seem to have an issue handling the amount if information per sample. Thresholding the color-space seems to preform marginally better than the extended clip length experiments, but are still not as as the experiments using longer strides. Despite the poor results of the threholding approach, all methods beat the ZeroR baseline method. Although the results may not be good enough to be deployed into a clinical setting, it shows that deep neural networks are a promising tool within the field of automatic semen analysis.

Looking at the table for the prediction of morphology results (Table 2), we see that pretty much all experiments lie around the ZeroR baseline. Most, however, beat the baseline by a small margin. It hard to make any strong conclusions about which methods work best, but it seems like using transfer learning for the stride experiments achieves better results than those trained from scratch. As for using different clip lengths, all methods seem to achieve a similar results. Overall, the results show that a more specific approach to predicting sperm morphology is needed, for example, analyzing individual spermatozoon using higher image resolutions.

4 CONCLUSION

In this paper, we presented the work done as part of the Medico Multimedia Task where we participated in two of the three available tasks. Overall, the results are promising and shows that neural networks are able to predict both motiltiy and morphology with a relatively low error margin. For future work, we aim apply 3D CNNs and more advanced architectures which may show an improvement over the presented results, in addition to exploring more advanced data preprocessing methods such as optical flow.

Method	Fold 1		Fold 2		Fold 3		Average	
	MA*E*	RMS*E*	MA*E*	RMS*E*	MA*E*	RMS*E*	MA*E*	RMS*E*
Stride Experiments								
Stride 1	10.436	14.769	11.079	15.155	11.581	14.404	11.032	14.776
Stride 5	8.563	11.856	9.843	13.754	12.172	15.510	10.192	13.707
Stride 10	9.358	12.711	9.477	15.524	11.892	15.065	10.242	14.433
Stride 25	9.490	13.530	7.149	9.579	10.871	14.532	9.170	12.547
Stride 50	10.005	13.961	9.804	14.468	10.691	13.593	10.167	14.007
TF Stride 1	9.874	13.408	8.450	11.638	10.257	13.972	9.527	13.006
TF Stride 5	10.937	14.699	7.903	10.544	10.322	13.217	9.721	12.820
TF Stride 10	8.714	11.955	8.256	11.153	9.917	13.029	8.962	12.046
TF Stride 25	8.505	11.211	8.818	11.889	10.480	13.919	9.268	12.340
TF Stride 50	9.021	11.505	9.604	11.943	11.338	14.818	9.988	12.755
Clip Length Experiments								
Clip Length 10	12.400	17.822	11.045	14.110	12.635	16.066	12.027	15.999
Clip Length 20	11.605	16.674	12.867	16.361	11.712	14.778	12.061	15.938
Clip Length 30	10.757	14.871	12.116	21.117	16.435	22.337	13.102	19.442
Clip Length 40	11.225	14.897	9.725	12.866	11.736	15.135	10.895	14.299
Clip Length 50	10.763	14.640	9.843	14.154	11.051	13.728	10.552	14.174
Threshold Experiments								
Stride 1	9.846	14.397	9.575	13.183	11.371	14.784	10.264	14.121
Stride 5	10.424	14.452	9.991	13.368	9.912	12.942	10.109	13.587
Stride 10	9.544	13.549	11.381	15.570	10.113	13.176	10.346	14.098
Stride 25	9.378	13.536	10.055	13.480	11.062	14.481	10.165	13.832
Stride 50	9.621	13.270	9.331	12.240	11.917	15.083	10.290	13.531
Baseline								
ZeroR	13.880	18.680	13.590	16.980	12.090	14.680	13.190	16.860

Table 1: The results for the *prediction of motility task*. Each entry shows the mean absolute error and root mean squared error for each fold of the three-fold cross-validation in addition to the average error across all folds.

Method	Fold 1		Fold 2		Fold 3		Average	
	MA*E*	RMS*E*	MA*E*	RMS*E*	MA*E*	RMS*E*	MA*E*	RMS*E*
Stride Experiments								
Stride 1	6.517	9.097	5.407	8.305	5.499	7.385	5.808	8.262
Stride 5	6.056	8.425	5.706	8.800	5.328	8.114	5.697	8.446
Stride 10	6.124	8.633	5.388	7.747	5.414	7.869	5.642	8.083
Stride 25	5.983	8.099	5.380	8.294	5.476	7.736	5.613	8.043
Stride 50	5.736	7.994	5.716	8.698	5.473	7.776	5.641	8.156
TF Stride 1	5.724	8.000	5.323	8.229	5.023	7.011	5.357	7.747
TF Stride 5	5.661	7.789	5.088	8.092	4.769	6.472	5.172	7.451
TF Stride 10	6.515	8.205	5.620	8.125	4.880	6.824	5.672	7.718
TF Stride 25	5.879	8.405	5.104	8.220	4.927	7.123	5.303	7.916
TF Stride 50	6.224	8.200	5.981	8.231	4.749	6.610	5.652	7.680
Clip Length Experiments								
Clip Length 10	6.216	8.793	5.636	7.899	5.295	7.634	5.716	8.109
Clip Length 20	6.336	8.355	5.604	7.753	5.112	7.241	5.684	7.783
Clip Length 30	6.097	8.485	6.342	9.315	5.666	8.177	6.035	8.659
Clip Length 40	6.059	8.645	5.744	8.665	5.122	7.501	5.642	8.270
Clip Length 50	6.211	8.794	5.584	8.677	5.282	7.946	5.692	8.472
Baseline								
ZeroR	5.990	7.950	5.990	8.270	5.820	8.130	5.930	8.100

Table 2: The results for the *prediction of morphology task*. Each entry shows the mean absolute error and root mean squared error for each fold of the three-fold cross-validation in addition to the average error across all folds.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] François Chollet and others. 2015. Keras. https://keras.io. (2015).

[3] Timothy Dozat. 2015. Incorporating Nesterov Momentum into adam.

[4] Trine B. Haugen, Steven A. Hicks, Jorunn M. Andersen, Oliwia Witczak, Hugo L. Hammer, Rune Borgli, Pål Halvorsen, and Michael A. Riegler. 2019. VISEM: A Multimodal Video Dataset of Human Spermatozoa. In *Proceedings of the ACM on Multimedia Systems Conference (MMSYS)*. https://doi.org/10.1145/3304109.3325814

[5] Steven Hicks, Pål Halvorsen, Trine B Haugen, Jorunn M Andersen, Oliwia Witczak, Konstantin Pogorelov, Hugo L Hammer, Duc-Tien Dang-Nguyen, Mathias Lux, and Michael Riegler. 2019. Medico Multimedia Task at MediaEval 2019. In *CEUR Workshop Proceedings - Multimedia Benchmark Workshop (MediaEval)*.

[6] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas de Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.