Stacked Dense Optical Flows and Dropout Layers to Predict Sperm Motility and Morphology

Vajira Thambawita[1,2], Pål Halvorsen[1,2], Hugo Hammer[1,2], Michael Riegler[1,3], Trine B. Haugen[2]

[1]SimulaMet, Norway   [2]Oslo Metropolitan University, Norway   [3]Kristiania University College, Norway

Contact:vajira@simula.no

ABSTRACT

In this paper, we analyse two deep learning methods to predict sperm motility and sperm morphology from sperm videos. We use two different inputs: stacked pure frames of videos and dense optical flows of video frames. To solve this regression task of predicting motility and morphology, stacked dense optical flows and extracted original frames from sperm videos were used with the modified state of the art convolution neural networks. For modifications of the selected models, we have introduced an additional multi-layer perceptron to overcome the problem of over-fitting. The method which had an additional multi-layer perceptron with dropout layers, shows the best results when the inputs consist of both dense optical flows and an original frame of videos.

1 INTRODUCTION

Our main goal of this task is to predict the sperm motility and sperm morphology from videos of sperm samples. In the 2019 Medico task [8], a video dataset was provided with ground truth values of sperm motility such as progressive motility, non-progressive motility, and immotility, and sperm morphology such as head defects, tail defects, and midpiece and neck defects. This task was introduced as completely new this year, and therefore, we could not find any previous work in previous mediaeval Medico task competitions [14, 15]. In this competition, the VISEM dataset [6] which contains sperm videos recorded from 85 participants is used. In the dataset paper, the authors presented baseline mean absolute error values for motility and morphology. Moreover, the importance of computer-aided sperm analysis can be identified from the research works which have been done to develop automatic sperm analysis method in last few decades [3, 13, 19].

Video analysis is a hot research topic in the field of deep learning. Some researchers are experimenting with video classification [2], detection [1], segmentation [5], and generations [12, 18] for various type of video datasets. Yue-Hei Ng et al. [20] experimented with video classification problem using well knows datasets such as sports-1M [10] and UCF101 [16]. In these experiments, they have generated dense optical flow images and row frames of videos to classify 120 seconds long videos. In this paper, we use very short video segments such as nine frames compared to these long segments such as 120s X 30 frames/s.

To solve this new regression problem of predicting morphology and motility from videos of sperm samples, this paper presents two deep learning methods where we used extracted dense optical flows and raw frames from the videos. In Section 2, we are going
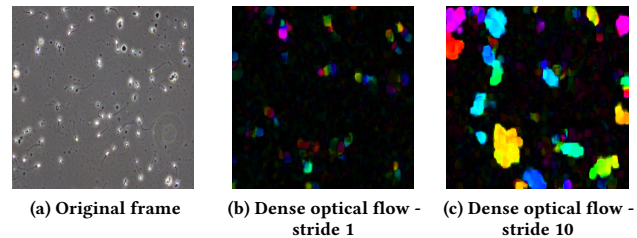
(a) Original frame (b) Dense optical flow - stride 1 (c) Dense optical flow - stride 10

Figure 1: Sample images used to construct input image stacks into the models

to present our two types of input data and two types of methods used in our experiments. Then, the results collected from these experiments will be discussed in Section 3. Finally, the paper ends up with conclusions and future work in Section 4.

2 APPROACH

We have selected the pre-trained ResNet-34 [7] to do some basic experiments of predicting sperm motility and sperm morphology using stacked normal raw video frames and a combination of stacked dense optical flows and raw frames of videos. In this paper, we obtain experimental results using two different types of inputs and from two different types of models.

2.1 Preprocessing data

To find estimates for the sperm motility and sperm morphology, we first preprocessed the input videos to generate two types of input. In the first type **(dataset - D1)**, we stacked nine consecutive frames from a video to make a single input data point. A sample of a raw frame of a video is given in Figure 1a. Before stacking raw video frames, we converted the RGB format frames of the video into grayscale images and resized them into 256x256. These nine frames represent nine different consecutive frames of a video. Moreover, we collected 250 stacked data points (chunks) from 250 locations in time from a video as described above.

For the second type of input **(dataset - D2)**, we generated a tensor with nine channels, which consists of a three-channels (RGB) original video frame (Figure 1a), a three-channels dense optical flow image of stride 1 (Figure 1b), and a three-channels image of dense optical flow of stride 10 (Figure 1c). The dense optical flow image of stride 1 was generated from two consecutive video frames from a selected location of a video. Then, we generated the stride-10 dense optical flow image using two frames; the first frame of the video chuck and the 10th frame of a selected video chunk. To generate dense optical flows [4] of two different frames of a video, the OpenCV library [9] was used with its inbuilt functions.
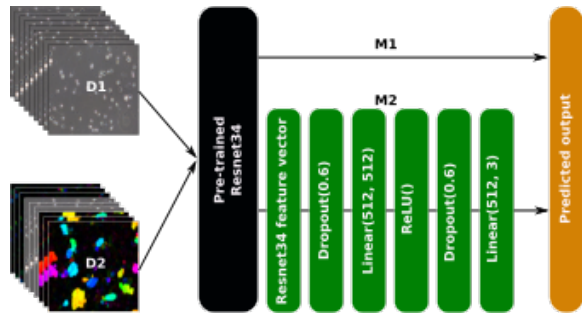
Figure 2: Big picture of our deep learning model: M1 - the base model of Resnet-34 with a three output last layer, M2 - the modified version of Resnet-34 with an additional MLP, D1 and D2 represent the two different types of input used in our experiments.

For both input types, we split the datasets into three folds based on the folds given in the video dataset provided by organizers. Then, a three-fold cross-validation was performed to evaluate our deep learning models which will be introduced in the later sections.

2.2 Deep learning model implementation

For implementation of our deep learning models, we selected Resnet-34 which is larger than the smallest, Resnet-18, and smaller than other large scales Resnet models like Resnet-50, Resnet-101, and Resnet-152. The selections of this intermediate Resnet-34 was done based on expandability of the model by adding additional multi-layer perceptron (MLP) within the available hardware resources (considering memory limitations of the available graphics processing units). In addition to that, the pre-experiments were done to identify over-fitting problems of strong models for simpler predictions and computation time required to finish training. Furthermore, expandability of the number of input channels of the model within the available GPU memory was examined.

For **method 1 (M1)**, we modified the input layer of the selected pre-trained Resnet-34 to take nine channel inputs and modified the last layer of the model to output only three values which are representing either three values of sperm motility or three values of sperm morphology. We used this method as our base model with the two different datasets (D1 and D2) as introduced in Section 2.1 and recorded results collected from this experiment in D1-M1 and D2-M1 rows in Table 1.

In **method 2 (M2)**, to avoid over-fitting problems of this task, we have embedded additional MLP to the end of the network with dropout layers [17]. The full structure of this additional MLP is depicted in Figure 2 using a green colour. The dropout values of this MLP were selected using pre-experiments, and it is a hyper-parameter for this model. The collected results of this method are tabulated in rows D1-M2 and D2-M2 of Table 1.

In the training process of all the above methods, the Adam optimizer [11] with a learning rate 0.001 was used. The mean square error (MSE) was used as the loss function for back-propagating error, and mean absolute error (MAE) was used for calculating the actual loss of predictions based on ground truth values of motility and morphology.

Table 1: MAE values collected from the proposed methods: D1-stacked gray-scale nine consecutive frames, D2-stacked an original frame + a dense optical flow image from two consecutive frames + a dense optical flow from two frames with stride=10; M1 - the basic model of Resnet-34 with modifications of number of input channels and outputs, M2 - the modified model with an additional MLP with dropout layers

Input	Method	Fold	Motility		Morphology	
			MAE	Average	MAE	Average
D1	M1	Fold 1	9.562		5.626	
		Fold 2	8.959	9.200	5.749	5.649
		Fold 3	9.079		5.573	
	M2	Fold 1	9.585		5.424	
		Fold 2	9.28	9.185	5.382	5.394
		Fold 3	8.689		5.375	
D2	M1	Fold 1	9.044		5.933	
		Fold 2	8.062	9.372	5.394	5.525
		Fold 3	11.01		5.248	
	M2	Fold 1	8.612		5.549	
		Fold 2	7.873	**8.825**	5.463	**5.293**
		Fold 3	9.991		4.868	

3 RESULTS AND ANALYSIS

According to the average MAE values shown in Table 1, the M2 method with the input type 2 (D2) shows best results among other methods and other input types. This method shows the best MAE value of 8.825 for the sperm motility and 5.293 for the sperm morphology. This improvement of error values can be seen as results of accumulated benefits of showing pre-processed temporal information such as dense optical flows to the model and the additional MLP to overcome the problem of over-fitting. Moreover, the added MLP in M2 gives better results with both input types (D1 and D2) for both predictions: sperm motility and sperm morphology. We achieved this performance as a result of the pre-processed input data with dense optical flows and the MLP introduced to overcome the over-fitting problem.

4 CONCLUSION AND FUTURE WORK

The input with a raw frame and dense optical flows of two difference stride values show better results compared to the stacked normal frames of videos. Moreover, the modified Resnet-34 model with an MLP which consists of dropout layers with high probabilities did achieve better results than the base model in the both cases because it helped to overcome the problem of over-fitting in the training stage. Finally, the combination of the input with dense optical flows and the modified Resnet-34 with an additional MLP shows the best overall performance.

In future work, it is worth to try CNN models with long short-term memory units to capture temporal features of video frames. Moreover, a 3D CNN can be a promising approach for this kind of task because 3D CNN models have capabilities to capture temporal information of videos.

REFERENCES

[1] Alan C Bovik. 2010. *Handbook of image and video processing*. Academic press.

[2] Darin Brezeale and Diane J Cook. 2008. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 3 (2008), 416–430.

[3] Karan Dewan, Tathagato Rai Dastidar, and Maroof Ahmad. 2018. Estimation of Sperm Concentration and Total Motility From Microscopic Videos of Human Semen Samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[4] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian conference on Image analysis*. Springer, 363–370.

[5] Arun Hampapur, Terry Weymouth, and Ramesh Jain. 1994. Digital video segmentation. In *Proceedings of the second ACM international conference on Multimedia*. ACM, 357–364.

[6] 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 10th ACM on Multimedia Systems Conference (MMSys'19)*. ACM, New York, NY, USA. https://doi.org/10.1145/3304109.3325814

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

[8] 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)*.

[9] Itseez 2014. *The OpenCV Reference Manual* (2.4.9.0 ed.). Itseez.

[10] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[13] Sharon T Mortimer, Gerhard van der Horst, and David Mortimer. 2015. The future of computer-aided sperm analysis. *Asian journal of andrology* 17, 4 (2015), 545.

[14] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Steven Alexander Hicks, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, Olga Ostroukhova, and Thomas de Lange. 2018. Medico Multimedia Task at MediaEval 2018.. In *Proceedings of the CEUR Workshop on Multimedia Benchmark Workshop (MediaEval)*.

[15] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas Lange, Kristin Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, Mathias Lux, and others. 2017. Multimedia for medicine: the medico task at MediaEval 2017. (2017).

[16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[17] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[18] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1526–1535.

[19] L. F. Urbano, P. Masson, M. VerMilyea, and M. Kam. 2017. Automatic Tracking and Motility Analysis of Human Sperm in Time-Lapse Images. *IEEE Transactions on Medical Imaging* 36, 3 (March 2017), 792–801. https://doi.org/10.1109/TMI.2016.2630720

[20] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4694–4702.