PREDICTING EMBRYO MORPHOKINETICS IN VIDEOS WITH LATE FUSION NETS & DYNAMIC DECODERS

Nathan H Ng & Julian McAuley
University of California, San Diego
{nhng, jmcauley}@ucsd.edu

Julian A Gingold & Nina Desai
Cleveland Clinic Foundation
{jgingold, ndesai}@ccf.org

Zachary C Lipton
Carnegie Mellon University
zlipton@cmu.edu

ABSTRACT

Annotating time lapse videos of developing embryos with their corresponding morphokinetic parameters is time-consuming and requires an experienced embryologist. We propose a 50-layer residual convolutional neural network, trained on a dataset of videos extracted from EmbryoScope incubators at the Cleveland Clinic Foundation. Specifically, we use late fusion to incorporate temporal context and a dynamic decoder to enforce monotonicity constraints in post processing. On a reduced label set of initial developmental stages, we demonstrate 87% per frame accuracy

1 INTRODUCTION

Historically, morphological evaluation of embryos has been the method of choice to select embryos for in vitro fertilization (IVF) (Cetinkaya and Kahraman, 2016). However, recent advancements in EmbryoScope incubators with built-in time lapse monitoring have allowed for the use of morphokinetic parameters to non-invasively determine embryo competence (Chamayou et al., 2013). These incubators use a high powered microscope and camera built into the device to take pictures on a cycle every 15 minutes.

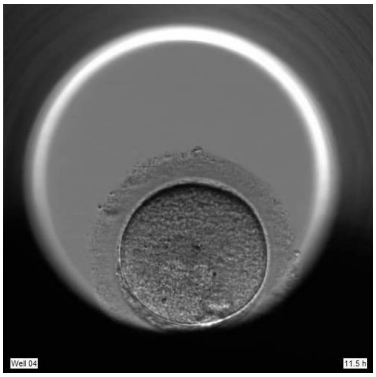
Current approaches to determining morphokinetic parameters from the resulting time lapse depend on a visual embryo assessment, typically by an experienced embryologist. Embryologists must view the sequence of photographs and manually annotate the time stamps at which each embryo achieves several developmental milestones. These scores are combined according to a heuristic formula in order to rank the embryos according to their (believed) viability for transfer into a prepared endometrium.

We investigate different methods of incorporating the temporal context of the video into the model architecture, as well as methods to enforce monotonicity constraints.

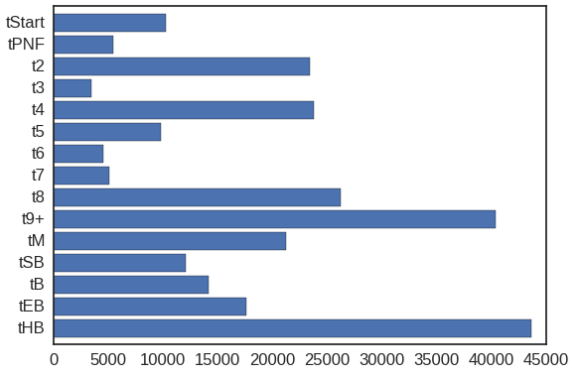
Our dataset consists of 1309 EmbryoScope time lapse videos extracted from incubators at the Cleveland Clinic Foundation. Each frame of the videos is a grayscale 500×500 resolution image with a well number in the lower left corner and an hour time marker in the lower right corner, as seen in Figure 1a. The videos in our dataset span 113 different patients, each with up to 12 wells and corresponding videos. Videos begin around 18 hours after fertilization, and end around 140 hours after fertilization.

For each video, we have manually labeled times at which the embryo was first observed in each developmental stage. We associate a stage label to each frame of video by examining the most recently assigned stage for the given well and patient. These stages are monotonically increasing, meaning that we cannot progress backwards once we have reached a certain stage.

For a large portion of our videos, we are missing expert ground truth annotations for the latter 9 stages. We thus restrict our dataset to the first 6 stages of development, cutting off each video at hour 70 to avoid class imbalance. This leaves us with 6 stages, including the initial stage (`tStart`),



(a) Sample frame from an EmbryoScope (not in dataset)



(b) Distribution of frames in each stage

the appearance and breakdown of the male and female pronucleus (`tPNf`), and the appearence of 2 through 4+ cells (`t2, t3, t4, t4+`). After processing, the frames are distributed 10.3% , 5.3%, 19.4%, 4.5%, 19.8%, and 40.7% in each stage respectively. We then split the dataset into training/validation/test sets by randomly selecting 93/10/10 patients and their respective wells. This gives us a total of 191449 / 21798 / 23348 frames respectively in each set for a percentage split of 81%/9%/10%.

2 METHODS

The embryo morphokinetics prediction task is a classification problem, where the input is an EmbryoScope video, and the output is a label indicating which stage the embryo is currently in at each frame of the video. Since an embryo's stage at any given point in time depends not only on... We investigate methods of combining frames in order extend our network to learn temporal features.

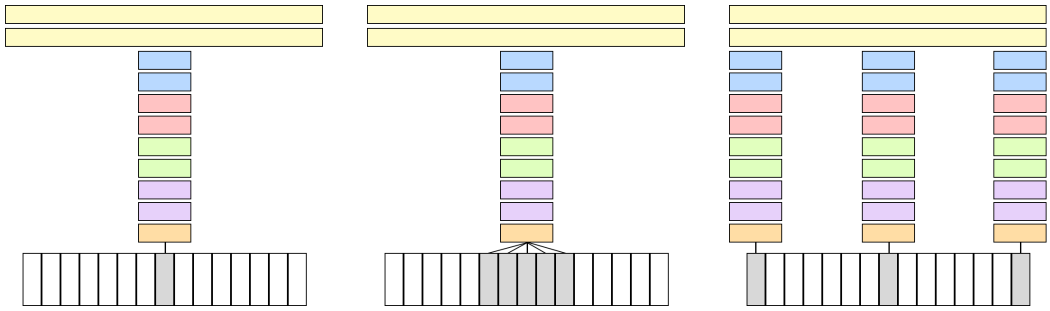
We first consider a baseline of a single frame architecture to understand how static images contribute to the per frame accuracy. We then consider two methods of temporal fusion: *early fusion* and *late fusion* (Karpathy et al., 2014). In early fusion, data is combined temporally at a pixel level. This is accomplished by stacking surrounding frames prior to applying any convolutions and modifying the number of channels in the input of our single framed model. In late fusion, data is combined temporally after the convolutional layers have been applied by concatenating the resulting fully connected layers. The same single frame model is applied on 3 separate frames: one in the past, one in the present, and one in the future.

Stages are monotonically increasing, meaning once we reach a certain stage we cannot go backwards. Our raw unstructured predictions do not have this constraint, so we enforce it in postprocessing by applying a dynamic decoder. We can define the error at any single frame for making a prediction of stage s given a predicted probability distribution p_i for observing stage i as some error function $e(s, p)$. We then seek to minimize this error across all frames of the video with the constraint that once we have transitioned into a later stage, we can no longer go back to an earlier stage.

We investigate two error functions, negative likelihood and earthmover distance. We define negative likelihood as $e(s, p) = -\log(p_s)$ and earthmover distance as $e(s, p) = |s - E(p)|$. We find that in all cases negative likelihood gives higher per frame accuracy and lower transition MAE and RMSE values.

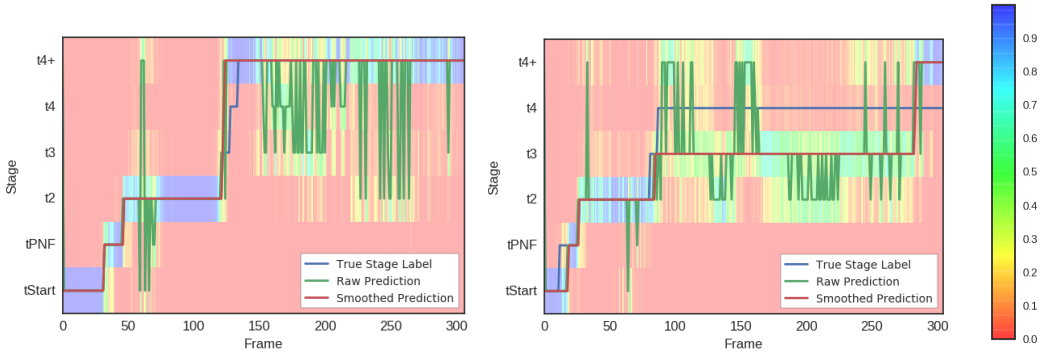
3 EXPERIMENTS

We now present our basic experimental setup. For all experiments we use the same 81%/9%/10% training/validation/test split. We report the per frame accuracy of our raw predictions, as well as the per frame accuracy of our smoothed predictions. We also report the mean absolute error (MAE) and root mean squared error (RMSE) in the frame timing of predicted smoothed transitions.



(a) Single Frame (b) Early Fusion (c) Late Fusion

Figure 2



(a) Helpful Smoothing (b) Unhelpful Smoothing

Figure 3

For all convolutional models we use a ResNet50 architecture pretrained on ImageNet(He et al., 2016). For single and early fusion models we modify the number of input channels to take in varying amounts of temporal context. In our late fusion model we concatenate the flattened fully connected layers together after applying the same convolutions separately. We also append the timestamp as a feature at the fully connected layer.

All models were trained using stochastic gradient descent with weight decay. We augment our dataset by applying random rotations and flips to images. Model weights are updated on the training set and we choose all non-differentiable hyper-parameters and architecture details based on validation set performance.

4 Related Work

The problem of predicting morphokinetic parameters for embryos has been addressed to some degree in Khan et al.. In that work they use an 8-layer convolutional network to count the number of embryos in an image. To incorporate temporal information they use

5 Discussion

References

Caroline Pirkevi Cetinkaya and Semra Kahraman. Morphokinetics of embryos - where are we now? *Journal of Reproductive Biotechnology and Fertility*, pages 1–8, 2016.

Sandrine Chamayou, Pasquale Patrizio, Giorgia Storaci, Venera Tomaselli, Carmelita Alecci, Carmen Ragolia, Claudia Crescenzo, and Antonino Guglielmino. The use of morphokinetic param-

Model	Raw Accuracy	Smoothed Accuracy	Transition MAE	Transition RMSE
ResNet50	0.8200	0.8460	11.225	29.650
Early Fusion (3 frames)	0.8237	0.8448	10.555	27.370
Early Fusion (9 frames)	0.8252	0.8423	10.927	29.400
Early Fusion (15 frames)	0.8182	0.8456	10.935	27.719
Early Fusion (9 frames + time)	0.8343	0.8430	11.152	26.321
Early Fusion (15 frames + time)	0.8420	0.8446	10.761	26.849
Late Fusion (15 frames)	**0.8479**	**0.8676**	**8.963**	**24.756**

Table 1: Results

eters to select all embryos with full capacity to implant. *Assisted Reproductive Genetics*, pages 703–710, 2013.

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*, pages 770–778, 2016.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

Aisha Khan, Stephen Gould, and Mathieu Salzmann. Deep convolutional neural networks for human embryonic cell counting.