# Automated embryo stage classification in time-lapse microscopy video of early human embryo development

Yu Wang, Farshid Moussavi, and Peter Lorenzen

Auxogyn, Inc.
1490 O'Brien Drive, Suite A, Menlo Park, CA 94025, USA
{ywang,fmoussavi,plorenzen}@auxogyn.com

**Abstract.** The accurate and automated measuring of durations of certain human embryo stages is important to assess embryo viability and predict its clinical outcomes in *in vitro* fertilization (IVF). In this work, we present a multi-level embryo stage classification method to identify the number of cells at every time point of a time-lapse microscopy video of early human embryo development. The proposed method employs a rich set of hand-crafted and automatically learned embryo features for classification and avoids explicit segmentation or tracking of individual embryo cells. It was quantitatively evaluated using a total of 389 human embryo videos, resulting in a 87.92% overall embryo stage classification accuracy.

**Keywords:** embryo stage classification, adaboost, bag of features

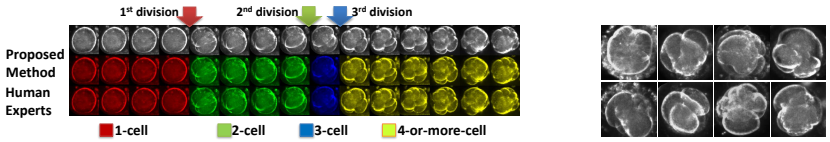
## 1 Introduction

There has been a rapidly growing need and interest in automated tools for assessing embryo viability quantitatively and objectively via non-invasive time-lapse imaging of early human embryo development. Timing parameters measured from time-lapse imaging of human embryos, such as the durations of 2-cell stage and 3-cell stage, have been confirmed to be correlated with the quality of human embryos and therefore can be used to select embryos with high developmental competence for transfer to *in vitro* fertilization (IVF) patients [1].

Accurately and objectively measuring these timing parameters requires an automated algorithm that can identify the stages of human embryo (i.e. number of cells) during a time-lapse imaging process (Figure 1(a)). In this work, we are interested in classifying human embryos into four stages, i.e. 1-cell, 2-cell, 3-cell, and 4-or-more-cell. This problem is challenging due to poor morphology of the embryos (Figure 1(b)), occlusion, and imaging limitations. The slight difference in appearance between human embryos after 1-cell stage imposes the greatest challenge for an automated algorithm.

Related studies in the literature exhibit difference in cell type, imaging modality and research objectives [2–6]. The works in [2][6] have focused on detecting

mitosis/cell division events. Although being an intuitive way to determine the stages of human embryo for our problem, explicit mitosis detection usually involves individual cell tracking and segmentation [2] which is very challenging to achieve [3]. Formulating the mitosis event detection as a classification problem in the spatial-temporal space [6] faces the problem of lack of positive training samples. The work in [4] employs the semi-markov model that uses intensity and gradient features extracted from temporal segments for HeLa cell cycle phase recognition in 3D time-lapse image sequences. Unlike the phases of HeLa cells that can be discerned based on intensity and shape, human embryo stages are generally more challenging to recognize and therefore require more powerful features and classification framework.




(a) (b)

**Fig. 1.** (a) An example of dark-field human embryo image sequence (sample frames) annotated by human experts and classified by proposed method. (b) Examples of challenging embryos due to poor morphology (e.g. fragmentation) and occlusion.

In this work, we present a 3-level classification method to classify embryo stages in time-lapse microscopy video. Our method avoids explicit segmentation and tracking of individual embryo cells. The classification method and embryo features can be easily adapted to various imaging modalities or even other cell classification and mitosis detection problems.

## 2 Methodology

Figure 2 illustrates the workflow of proposed embryo stage classification method. Given a human embryo video, 62 standard hand-crafted features and 200 automatically learned bag-of-features (BoF) are extracted from each frame of the video. A level-1 Adaboost classification model performs classification on each frame independently. It consists of 4 Adaboost classifiers, each trained for classifying one class from the rest classes using the 262 features. A level-2 Adaboost classification model consists of another 4 Adaboost classifiers trained with augmented feature set that includes additional features computed from the level-1 classification probabilities in a local temporal window. Level-2 Adaboost is designed to exploit local temporal context and refine the level-1 classification result. At level 3, the Viterbi algorithm takes as input the level-2 classification probabilities and a temporal image similarity measure and generates the final embryo stage classification result within a global temporal context.

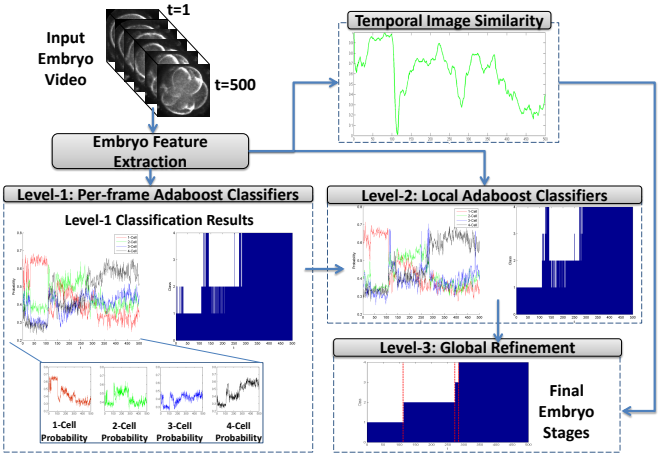



**Fig. 2.** Flowchart of the proposed method. Red dotted vertical lines in the level-3 plot indicate the ground-truth of stage transition points.

### 2.1 Embryo Features

Our embryo features include 62 hand-crafted features (22 Gray-Level Co-occurrence Matrices (GLCM) features, 10 Local Binary Patterns (LBP) features [7], 15 Hessian features, 10 Gabor features, and 5 region properties) and 200 bag-of-features [8]. The GLCM, LBP, and Gabor features are well-known texture features. Hessian features are statistics (e.g. mean intensity value) computed from Hessian-filtered embryo images in which the cell edges are enhanced. The region properties (area, number of convex hull points, solidity, eccentricity, and perimeter) are computed from a rough embryo mask obtained by applying a shortest path algorithm to extract the embryo boundary in polar image space.

Bag-of-features are based on keypoint descriptors such as SIFT [9]. Both densely sampled descriptors and sparsely detected descriptors (Figure 3(a)) are used in the method. $K$-means clustering was applied on sparse and dense SIFT descriptors computed on training embryo images respectively, resulting in a codebook with 200 codewords which are the centroids of clusters. Given a testing image, sparse and dense descriptors are extracted first and then quantized by hard-assigning each descriptor to one codeword. The final BoF (Figure 3) is an occurrence histogram that represents the frequency of the codewords.

The additional level-2 features are temporal features computed from the classification probabilities output by level-1 Adaboost. At each frame, statistics (mean, median, max, min, and standard deviation) of the classification probabilities of each class are computed from its local temporal window (e.g. 5 frames). Therefore a total of 20 level-2 features are added to the original feature set.

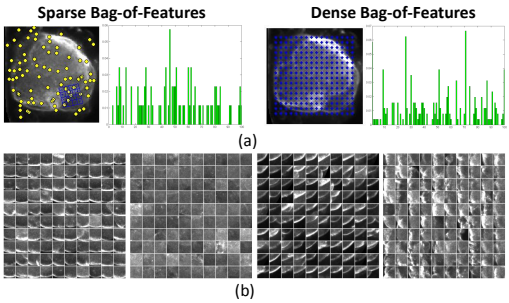



**Fig. 3.** Illustration of the Bag of Features. (a) Dense and sparse occurrence histograms generated from sparsely detected descriptors and densely sampled descriptors with learned codebook. (b) Four examples of clusters (appearance codewords) generated by $k$-means clustering.

### 2.2 2-Level Adaboost Classification Model

We employ the one-vs-all scheme to handle this multi-class classification problem with binary Adaboost classifiers. There are four Adaboost classifiers at each level of the 2-level Adaboost classification model. Each Adaboost classifier consists of $N$ decision stumps (one-level decision tree) and is trained to separate one class from the other classes. For an Adaboost classifier trained for class $i \in \{1, 2, 3, 4\}$, its output for a embryo image at $t$ is

$$P(y_t = i|x_t) = \frac{\sum_{k=1}^{N} a_{ik} h_{ik}(x_t)}{\sum_{k=1}^{N} a_{ik}}, \tag{1}$$

where $x_t$ is the extracted feature vector for the image, $a_{ik}$ is the weight of the base classifiers, $h_{ik} \in \{-1, 1\}$ is the output of the base classifiers, and $P(y_t = i|x_t)$ is the posterior classification probability normalized to $[0, 1]$ (Figure 2).

### 2.3 Temporal Image Similarity

Besides representing embryo images in proposed method, the BoF is also used to compute a temporal image similarity measure that is subsequently used by the Viterbi algorithm to define state transitional probability. Given the normalized BoF histograms of two consecutive embryo frames ($t - 1$ and $t$), the temporal image similarity $d(t)$ is defined based on the Bhattacharyya distance of the two histograms. This temporal similarity measure based on BoF is registration free since SIFT is rotation invariant and translation of the embryo has been corrected by cropping the image to a smaller patch centered on the center of embryo. One example of the temporal image similarity is shown in Figure 4. "Dips" in the plot are good indications of stage transitions.

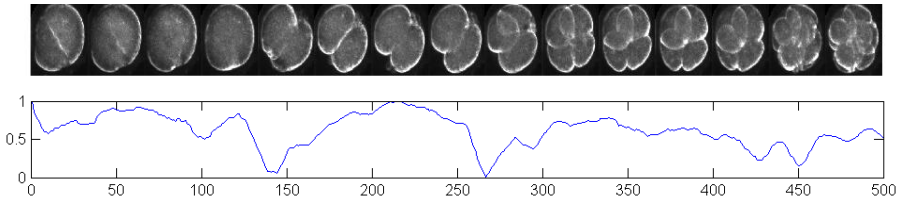



**Fig. 4.** An example of the temporal image similarity.

### 2.4 Global Embryo Stage Refinement

At level-3 of the proposed method, the Viterbi algorithm is employed to refine embryo stages within global temporal context. The problem is to infer the best state sequence of embryos that maximizes the posterior probability $P(Y|X)$:

$$\hat{Y} = \arg\max_{Y} P(Y|X), \tag{2}$$

where, $Y = \{y_1, ..., y_T\}$ is the state sequence, $X = \{x_1, ..., x_T\}$ are the feature vectors representing the embryo images.

The Viterbi algorithm recursively finds the weight $V_{t,i}$ of the most likely state sequence ending with each stage $i$ at time $t$.

$$V_{1,i} = P(x_1|y_1 = i)P(y_1 = i), \tag{3}$$
$$V_{t,i} = P(x_t|y_t = i) \max_{j}(P(y_t = i|y_{t-1} = j)V_{t-1,j}), t \neq 1. \tag{4}$$

where, $P(y_1 = i)$ represents the prior probability of each class at the first frame and is set to 1 for $i = 1$, $P(x_t|y_t = i)$ is the observation probability, and $P(y_t = i|y_{t-1} = j)$ is the transition probability. If we assume the 4 stages are equally probable after the first frame, the observation probability $P(x_t|y_t = i)$ is simply the classification probability output by the level-2 Adaboost. The transition probability $P(y_t = i|y_{t-1} = j)$ is defined as a frame-dependent stage transition matrix $A(t) = \begin{pmatrix} d(t) & 1-d(t) & 0 & 0 \\ 0 & d(t) & 1-d(t) & 0 \\ 0 & 0 & d(t) & 1-d(t) \\ 0 & 0 & 0 & 1 \end{pmatrix}$, where $d(t)$ is the temporal image similarity. When two consecutive frames are almost the same (i.e. $d(t)$ is close to 1), the transition matrix favors no embryo stage change.

## 3 Experimental Studies

We collected human embryo videos from a variety of clinical sites and evaluated the classification performance of our proposed method on them.

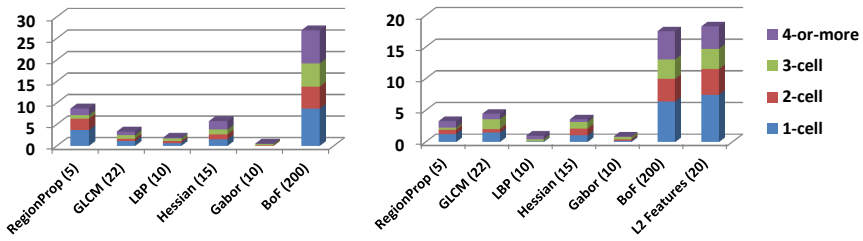



**Fig. 5.** Importance of different sets of features in trained level-1 (left) and level-2 (right) classification models.

**Table 1.** Classification performance at different levels.

| | 1-cell | 2-cell | 3-cell | 4-or-more | Overall |
|---|---|---|---|---|---|
| Level-1 | 87.96% | 77.45% | 7.79% | 85.03% | 80.10% |
| Level-2 | 88.04% | 72.05% | 10.71% | 92.94% | 82.53% |
| Level-3 | 91.95% | 85.58% | 20.86% | 94.14% | 87.92% |

### 3.1 Dataset and Ground Truth

All of our embryo videos were captured using the Eeva$^{TM}$ system[1] which comprises of four inverted digital microscopes modified for darkfield illumination. Embryo images were acquired every 5 minutes for up to 2 days until the majority of embryos reached the four-cell stage. The first 500 frames of each embryo video were kept for analysis and each frame was cropped to a size of $151 \times 151$ pixels from an original $451 \times 451$ pixels using the center of embryo. Two human experts annotated the embryo stages for 716 human embryo videos. Among these videos, 327 embryos were used for training and 389 embryos were used for testing. For the training data, embryo frames for which the two human experts disagreed about the embryo stage were excluded. Since the 3-cell stage is usually very short, we have far fewer 3-cell training samples than the other classes. The training data is split into two halves for the training of level-1 and level-2 Adaboost classifiers, respectively. For the testing data, the embryo stage ground-truth is the average of the annotations by the two human experts.

### 3.2 Evaluation Results

The proposed method is implemented in Matlab. On an average laptop the computation times for the extraction of 262 features, the computation of the temporal image similarity, and the 3-level classification of embryo stages per embryo sequence were approximately 53.2, 0.052, and 0.097 seconds. We quantified the importance of each of the 262 features by summing up the weights of

[1] `http://www.auxogyn.com/eeva.phpp`

the base stump classifiers that selected the feature in trained Adaboost classifiers. Figure 5 shows the importance of each feature set which is the sum of the importance of its features and indicates that BoF and level-2 features played the most important role in level-1 and level-2 classification models, respectively.

**Table 2.** Confusion matrix of the final classification result.

| | 1-cell | 2-cell | 3-cell | 4-or-more |
|---|---|---|---|---|
| 1-cell | 43276 (91.95%) | 3399 (7.22%) | 245 (0.52%) | 143 (0.3%) |
| 2-cell | 643 (1.31%) | 41866 (85.58%) | 2518 (5.15%) | 3891 (7.95%) |
| 3-cell | 5 (0.05%) | 4070 (43.36%) | 1958 (20.86%) | 3353 (35.72%) |
| 4-or-more | 0 (0%) | 2620 (2.94%) | 2603 (2.92%) | 83910 (94.14%) |

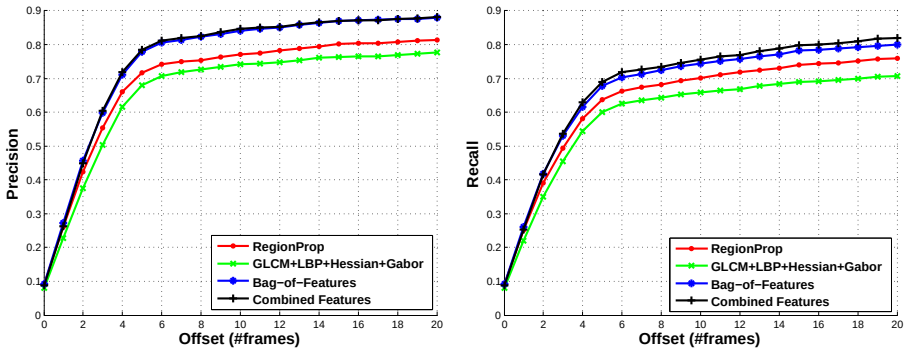
In the first experiment, the embryo stages predicted by proposed method are compared with ground-truth. The overall classification accuracy and classification accuracy of each class are shown for each level of the method at Table 1. The confusion matrix for the final classification results is shown in Table 2. Each level of the classification improves the overall classification accuracy over the previous level. Due to the lack of 3-cell training samples and their resemblance to 2-cell and 4-or-more-cell embryos, only 7.79% accuracy was reached by the level-1 Adaboost for 3-cell class. This accuracy was improved to 20.86% in the final level.

In the second experiment, the three division frames detected by classification were also compared with the ground-truth. A detected division frame is considered as a true-positive if it is within certain offset tolerance to the ground-truth and considered as a false-positive otherwise. A ground-truth division frame is considered as false-negative if there is no detected division frame within certain tolerance.

Figure 6 shows the precision and recall as functions of the offset tolerance. The precision and recall curves for three subsets of features were generated to evaluate their contributions to the final classification performance separately. It can be seen from Figure 6 that BoF outperformed the hand-crafted features and their combination (262 features) reached the highest performance.

## 4 Conclusion and Future Work

We presented a classification method for effectively classifying embryo stages in time-lapse microscopy of early human embryo development. When applied to a large testing dataset collected from multiple clinical sites, the proposed method achieved a total of 87.92% classification accuracy. In our future work, we plan to improve our embryo quality predictors using timing parameters extracted by this novel embryo stage classification model.




(a) (b)

**Fig. 6.** Precision (a) and Recall (b) of cell division detection as functions of the offset tolerance. The same 3-level Classification method is used for each set of the features.

## References

1. Wong, C., Loewke, K.E., Bossert, N.L., Behr, B., J De Jonge, C., Baer, T.M., Reijo Pera, R.A.: Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. Nature Biotechnology, vol. 28(10), pp. 1115–1121. (2010)
2. Yang, F., Mackey, M.A., Ianzini, F., Gallardo, G., Sonka, M.: Cell Segmentation, Tracking, and Mitosis Detection Using Temporal Context. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 302–309. (2005)
3. Huh, S., Ker, D.F., Bise, R., Chen, M., Kanade, T.: Automated mitosis detection of stem cell populations in phase-contrast microscopy images. IEEE Transactions on Medical Imaging, 30(3), pp. 586–596. (2011)
4. El-Labban, A., Zisserman, A., Toyoda, Y., Bird, A.W., Hyman, A.: Discriminative Semi-Markov Models for Automated Mitotic Phase Labelling. IEEE International Symposium on Biomedical Imaging (ISBI), pp. 760–763. (2012)
5. Harder, N., Mora-Bermdez, F., Godinez, W., Ellenberg, J., Eils, R., Rohr, K.: Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 840–848. (2006)
6. Li, K., Miller, E.D., Chen, M., Kanade, T., Weiss, L.E., Campbell, P.G.: Computer vision tracking of stemness. IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), pp. 847–850. (2009)
7. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), pp. 971–987. (2002)
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2169–2178. (2006)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), pp. 91-110. (2004)