2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Preliminary Study on Retro-reconstruction of Cell Fission Dynamic Process using Convolutional LSTM Neural Networks

Yuding Wang, Ting Gang Chew, and Liangjing Yang, *Member, IEEE*

Abstract— **Cell morphological analysis has great impact towards our understanding of cell biology. It is however technically challenging to acquire the complete process of cell cycles under microscope inspection. Using convolutional long short-term memory (ConvLSTM) networks, this paper proposes a comprehensive visualization method for cell cycles by retro-reconstruction of the preceding frames that are not captured. Results suggested that this method has the potential to overcome existing technical bottlenecks in image acquisition of cellular process and hence facilitate cell analysis.**

Clinical Relevance— **This model allows back-tracing to complete the visualization of the cellular processes through a short segment of microscope-acquired cellular changes hence providing a starting point for exploring applications in predicting or backtracking unknown cellular processes.**

I. Introduction

Understanding cell cycle regulation is of vital importance to our understanding of biological process including cancer development [1]. A complete life cycle of eukaryotic cells is normally divided into four phases: first gap phase (G1 phase), synthesis phase (S phase), mitosis phase (M phase), and second gap phase (G2 phase) [2]. However, no matter classifying cell cycle manually or identifying and isolating cells into a particular cell phase [3], these methods can hardly capture target cells with recording of the entire cell cycle, which are essential information in the visualization of cell morphology. For in vivo cell inspection, the current classification system can accurately identify the cell stage [4] and extract a cell at a particular phase. However, since living cells are changing all the time, the cells after extraction have changed compared to the initially observed cell state. The system does not know the previous state of the single cell, which can help complete the entire cycle of the cell. According to cell section analysis, the extraction of cell geometric information at a definite time lacks continuity. The retrospection of the cell's past state and prediction of the future potential state can help researchers better understand the geometric characteristics of the cell. Researchers in the field of cell morphology, at present, concentrated on the spatial characteristics of cells, and paid little attention to the spatiotemporal analysis of sequential information in cells, especially in the aspect of machine learning (ML) based methods.

In recent years, deep learning has been widely used in extracting cellular spatial information. Several outstanding models play notable roles in the analysis of cellular activities. U-net network is used to segment cell images [5]; deep CNN can accurately and rapidly classify leukocyte cell and red blood cell [6, 7]. There are also various methods having been researched for time-space sequence problem in other fields. LSTM, a variant of recurrent neural network (RNN), is widely applied to process time sequence by loops. It can memorize the past input to exploit information about sequence input using the loop structure [8]. Compared with LSTM. ConvLSTM model was developed to build an end-to- end trainable model for the precipitation nowcasting problem [9]. This model has better performance for extracting spatial information by adding convolution operations on the basis of LSTM. Meanwhile, based on CNN, 3D convolutional networks (3D CNN) can be utilized to capture the temporal information encoded in multiple contiguous frames [10].

This study model cellular spatiotemporal relationships using ConvLSTM followed by 3D CNN to decode the spatiotemporal information. This architecture is established to backtrack the cellular process so as to get previous states of the single cell. After that, we will use U-net model to extract the geometric features and evaluate the result [5].

Our contributions can be summarized as (A) an end-to-end trainable deep neural network model for completing the cell process; (B)a method to dynamically adjust the input so that the system can automatically use a model with limited trained data and fewer parameters to get missing frames. This system has the ability to complete the cell state in real-time. A brief introduction to our proposed model is presented in Section II. Section III presents the database and experiment settings. The experimental results are discussed and analyzed in Section IV. Finally, Section V concludes this paper by summarizing the contribution of the work and suggestionof potential future work.

II. PROPOSED MODEL

A. 3D CNN

In this study, we use 3D CNN as the decoder to upsample the reduced spatiotemporal feature. In 2D convolutional network, convolution focuses on spatial information at the expense of temporal information. However, as demonstrated in Fig.1, the 3D convolution and pooling operations use a 3D kernel to overlap cubes spatiotemporally, hence, retaining the timing information of the input signal to produce an output volume [11]. Compared with 2D CNN, 3D CNN have better

* This work was supported by the Zhejiang University/University of Illinois at Urbana-Champaign Institute, funded in part by Natural Science Foundation of China (61950410605) and Zhejiang Province (LQ19F030013) led by Principal Supervisor Liangjing Yang, and collaboratively supported by Joint Research Center for Adaptive, Resilient Cyber-Physical Manufacturing Networks, DREMES, ZJUI Institute.

Yuding Wang, Ting Gang Chew, and Liangjing Yang are with International Campus, Zhejiang University | ZJU-UIUC institute, Haining, Zhejiang; e-mail: {yuding.17; tinggchew@intl.zju.edu.cn; liangjingyang} @intl.zju.edu.cn)..

temporal information modeling capabilities as illustrated in Fig 1.

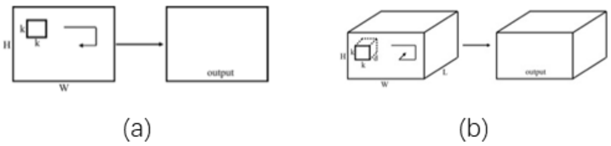

Fig. 1. 2D and 3D convolution operations [11]. a) applying 2D convolution on an image results in an image. b) applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

B. *ConvLSTM network*

While LSTM is capable of processing temporal data, if the temporal data is an image, it will be more effective for image feature extraction by adding convolution operations on the basis of LSTM, which has the capability of dealing with spatial information. Therefore, a new network combining convolution operations and LSTM is proposed, which is termed ConvLSTM [9]. The structure of ConvLSTM layers is illustrated in Fig. 2.

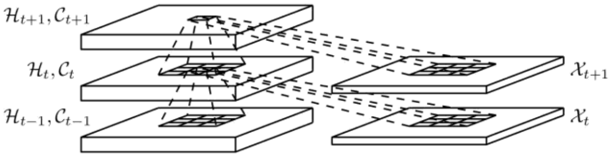
$\mathcal{H}_{t+1}, \mathcal{C}_{t+1}$
$\mathcal{H}_t, \mathcal{C}_t$
$\mathcal{H}_{t-1}, \mathcal{C}_{t-1}$
$\mathcal{X}_{t+1}$
$\mathcal{X}_t$

Fig. 2. Structure of a convolutional LSTM

Convolutional LSTM is invented to predict precipitation nowcasting from radar images, which could not fulfill our aim in completing the partially acquired video of the cellular process. Thus, we adopted a modified version of ConvLSTM, which is more sensitive to motions [12], formulated as follows:

$$i_t = \sigma(W_t^i \mathbf{I}_t + W_h^i * h_{t-1} + b^i) \tag{1}$$

$$f_t = \sigma(W_t^f \mathbf{I}_t + W_h^f * h_{t-1} + b^f) \tag{2}$$

$$\tilde{c}_t = \tanh\,(W_t^{\tilde{c}} \mathbf{I}_t + W_h^{\tilde{c}} * h_{t-1} + b^{\tilde{c}}) \tag{3}$$

$$c_t = \tilde{c}_t \odot i_t + c_{t-1} \odot f_t \tag{4}$$

$$o_t = \sigma(W_x^o \mathbf{I}_t + W_h^o * h_{t-1} + b^o) \tag{5}$$

$$h_t = o_t \odot \tanh\,(c_t) \tag{6}$$

where $\mathbf{I}_t$, h_t, and c_t represent the input tensor, the hidden states, and the cell states, respectively. The gate activations at current time step are represented by i_t, f_t and o_t. Notation "$\odot$" denotes the Hadamard product operation while the convolution operation is denoted by "$*$".

C. *U-net network*

In this study, U-net is utilized to extract the geometric features and evaluate the effects. U-net is a full convolution network, commonly used in cell segmentation [5]. There are many U-net network typessince Ronnberg*et al.* [5]described the basic concept of U-netas illustrated in Fig.3. As we are only using U-net to segment the outcomes and original images, we will not explore the difference of U-net variants. In this experiment, we are using the original U-net architecture with a slight customization of the depth and channels.

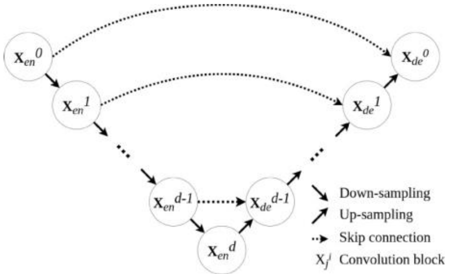

Fig. 3. U-Net architecture [5] .

III. Data Collection And Processing

A. *Dataset*

In this experiment, the dataset we used is the 246 x 233 image sequence of four different cells' complete living cell division process. Each process takes 80 x 5 frames. Before the experiment, we do primary data augmentation (center crops, rotations, hue, saturation, and exposure shifts) for four image sequences. Frames from a portion of the augmented image sequence are extracted and divided into a new image sequence. Therefore, we get 18 periods of video with 80 frames, 18 with 40 frames, and 24 with 20 frames (a total of 2540 training images). The test image is an unknown cellular process without beginning and ending. Subsequently, resizing of all test images into 128x128 pixels. A representative diagram of the process is shown in Fig.4.

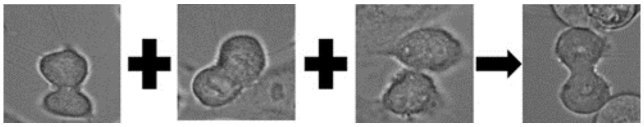

Fig. 4. Three images on the left are three samples of training image and the extreme right one is the sample of test image

B. *Experiment setting*

The architecture developed contains three ConvLSTM layers and one 3D convolution layer, which is illustrated in Fig. 5. For all experiments, the inputs were size Nx128x128, where N is the number of input frames and the output was of size 1x128x128. To further gather spatiotemporal features [13], the output layer is the 3D convolution layer with 3x3x3 kernel since the 3x3x3 convolution kernel is shown to have the best performance in a limited framework [11]. Also, the hidden layers of ConvLSTM is 128,128 and 64. For optimizer, we adopted the Adam optimizer [14] with adaptive learning rate and a batch size of 8.

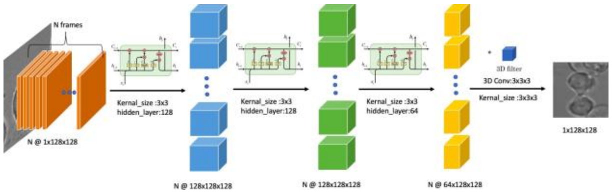

Fig. 5. Overview of completing cellular process system. This structure includes three ConvLSTM layers and one 3D convolution layer

IV. EXPERIMENTRESULTS

A. Backtracking the starting frame

As mentioned in the introduction, the critical concept for inferring the previous state is the same as predicting the subsequent frames. The main distinction is the order of test and training data. Therefore, the core of this experiment is to answer two questions [15, 16]: (1) How far can we predict/review? (2) How accurate can we get based on the given limited data source?

In question (1), the essential thread is to predict the following unknown frames based on groundtruth and the previous frame prediction. Since the first predicted image's quality has a bias with the groundtruth, the following image will inevitably be increasingly different from the original image. Both GANs scheme [17] and ConvLSTM scheme [9], cannot resolve this dilemma due to high dimensionality and complexity of natural images [16].

However, unlike the traditional video prediction task, our task is to complete the missing cellular change process, which means we have a relatively abundant known continuous image sequence. Therefore, we choose to deduce the unknown state image by using the frames with different intervals rather than using the traditional method of predicting motion frame by frame. For instance, suppose a complete cell division process is 30 frames long, and we have the 6th to 30th frame of them. Now we want to get the first 5 images. Because the current number of consecutive frames and the number of lost frames is changing in practice, we cannot set the dimensions of input and output dimensions in a targeted manner. To overcome this problem, we train a model using 6 frames with different intervals to predict one frame and develop a method to change the input dynamically for different targets, using the example above, if we want to get the third frame. The algorithm will automatically use image [6,9,12,15,18,21] as input to get a predicting output image. We use image [7,11,15,19,23,27] as input to get another predicting output image. The result is the average of two outputs. The predicted image based on different intervals are shown in Fig. 6.

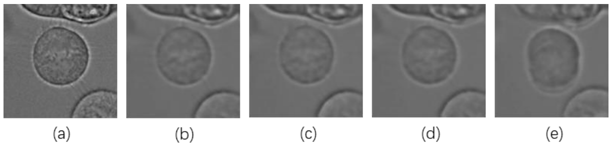

Fig. 6. (a) is the original image; (b),(c),(d) is the result produced by six frames with different interval; (e) is the 5th image directly backtracking by six frames

The figure shows that the outcome predicted from image sequence with five frames interval is much better than the outcome backtracking directly using ConvLSTM. To further evaluate the quality of results so as to answer the "How accurate", Question (2), we randomly choose 8 frames as the ground truth. For each ith target image, we use six frames with the specific interval to induce the cell image. After that, we compare the outputs with target images quantitatively and record the result in table 1. The quality of the backtracking frames is assessed by the average of intensity-based mean squared error (MSE) [15]and the average of structural similarity index (SSIM) [18]. Table 1 compared the performance of the backtracking image using the same number of frames with different frames interval. From the average SSIM, we can find that the structure similarity decreases as the number of interval frame increase, which means the result is blurred compared with ground truth. For this result, we speculate that when the cell is not dividing, the spatial information of the cell changes slowly and uniformly most of the time. Hence, the outputs of the image sequence with shorter intervals have high average quality. Although the outcomes are blurry compared to the original image, they can play a role in single cell analysis since we can still recognize the contour of the single cell.

TABLE I. MEAN MSE AND MEAN SSIM BETWEEN THE ORIGINAL FRAMES AND THE BACKTRACKING IMAGES BY DIFFERENT METHODS

	MSE	*SSIM*
Continuous Frames	105.6	0.81
3 Frames intervals	129.3	0.77
5 Frames intervals	132.4	0.74
5th Image directly backtracking by frames	263.4	0.55

B. Identifying the contour using U-net

To that the prove the backtracking images can be used to complete incomplete cell process, we use U-net to draw the contour of the original image and backtracking images, then compare the similarity and other characteristics of these images mentioned in Section 2. Compared with other mainstream contour recognition methods such as active contour models, U-net network has more robust processing capabilities for uneven grayscale images and can better recognize target contours after sufficient training [5]. Therefore, our experiment eliminates the interference of "fuzzy" factors as much as possible, so as to objectively evaluate whether the cell contour of the output picture is valuable as an auxiliary material for single cell analysis.

In Fig.7, we visually and intuitively compared the differences in contours. The first three pictures are outlines obtained by U-net network. Then we remove the black background of the picture and the disconnected area. Next, we overlap the picture and adjust the visual effect by setting the channel and transparency. Visually, from Fig.7(d), we observe that the outlines of the two overlaps are almost the same although Fig.7(b) are fuzzier than Fig.7(a), which means ConvLSTM captures the high-dimensional information of the outline well. As for Fig.7(e), due to the larger interval, the contour has a slight upward shift relative to the real contour in Fig.7(a). However, the area and circumference smoothness of the two are similar, which indicates that this outcome is still useful during cellular analysis. The result of the quantification is shown in Table 2. We quantitatively compare the contour of the original image with the contour of the image obtained by retro-reconstruction. The high MSE value suggested that the relative position of the contours from retro-reconstruction outcomes is shifted from the actual contour, while the high SSIM value proves the similarity of the contour shape.

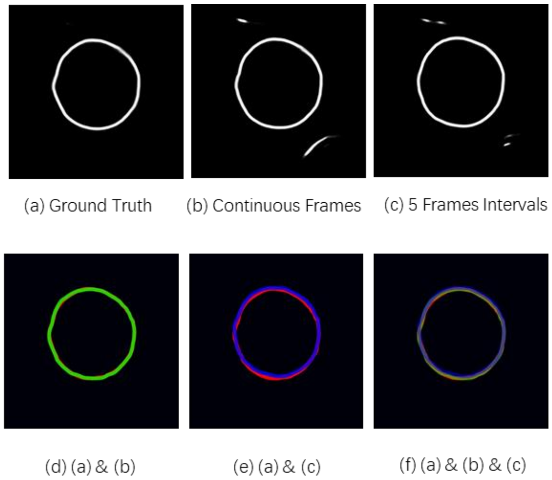

Fig. 7. Visual comparison; (a) original image; (b) 6 continuous frames of different intervals; (c) 6 frames with 5 frame intervals; (d), (e), (f) overlaps of results (blue/green) & original image(red)

TABLE II. MSE AND SSIM BETWEN THE ORIGNAL CONTOUR AND THE CONTOUR OF BACKTRACKING IMAGSE WITH DIFFERENT INTERVAL

	MSE	*SSIM*
Continuous frames	191.4	0.98
3 frames intervals	315.3	0.96
5 frames intervals	405.7	0.95

V. CONCLUSION

Our work combines 3D convolution and ConvLSTM model to propose a method for more comprehensive visualization of the cell cycle by retro-reconstruction of the uncaptured preceding frames. Using future frames with different intervals, this model can backtrack relatively distant states of cell without losing too much accuracy. Results suggest that this method may have the potential to overcome the existing technical bottlenecks in image acquisition of cellular processes and facilitates cell analysis.

It is worth noting that the current structure of the network still has much room for improvement. To be more specific, one possible limitation is that the MSE loss function we are using might be guiding the model to gather global features in the entire frames rather than the single cell, which may take some unnecessary information into consideration. Moreover, in this experiment, MSE may not be the best way to evaluate the performance. The next step would be developing corresponding preprocessing algorithms to emphasize cellular characteristics such as contour or other structural features, as well as exploring alternative objective functions to boost the performance.

Beyond the network structure, it is also very intriguing to explore the backtracking performance. We are also planning to predict the possibilities of cell differentiation using ConvLSTM or GANs structure, which may have implications for considerable practical applications, such as genetic engineering and pharmacological insights.

VI. ACKNOWLEDGMENTS

This work was supported by the Zhejiang University/University of Illinois at Urbana-Champaign Institute, funded in part by Natural Science Foundation of China (61950410605) and Zhejiang Province (LQ19F030013) led by Principal Supervisor Liangjing Yang, and collaboratively supported by Joint Research Center for Adaptive, Resilient Cyber-Physical Manufacturing Networks, DREMES, ZJUI Institute..

REFERENCES

[1] B. P. Ingalls, B. P. Duncker, D. R. Kim, and B. J. Mcconkey, "Systems level modeling of the cell cycle using budding yeast," *Cancer Inform*, vol. 3, 2007.

[2] Bo Yang Yu, Caglar Elbuken, Carolyn L. Ren, and Jan P. Huissoon, "Image processing and classification algorithm for yeast cell morphology in a microfluidic chip," *Journal of Biomedical Optics*, vol. 16, no. 6, pp. 066008, 2011.

[3] Shinsuke Ohnuki, Satoru Nogami, and Yoshikazu Ohya, "A microflu- idic device to acquire high-magnification microphotographs of yeast cells," *Cell Division, 4, 1(2009-03-24)*, vol. 4, no. 1, pp. 1–8, 2009.

[4] A. I. Shahin, Yanhui Guo, K. M. Amin, and Amr A. Sharawi, "White blood cells identification system based on convolutional deep neural learning networks," *Computer Methods & Programs in Biomedicine*, p. S016926071730411X, 2017.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convo- lutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[6] Laith Alzubaidi, Omran Al-Shamma, Mohammed A. Fadhel, Laith Farhan, and Jinglan Zhang, *Classification of Red Blood Cells in Sickle Cell Anemia Using Deep Convolutional Neural Network*, 2020.

[7] Mehdi Habibzadeh, Adam Krzyżak, and Thomas Fevens, "White blood cell differential counts using convolutional neural networks for low resolution images," In *International Conference on Artificial Intelligence and Soft Computing*, pp. 263-274. Springer, Berlin, Heidelberg, 2013.

[8] F. A. Gers, "Learning to forget: continual prediction with lstm," in *9th International Conference on Artificial Neural Networks: ICANN '99*, 1999.

[9] Xingjian Shi, Zhourong Chen, et al., "Convolutional lstm network: A machine learning approach for precipitation nowcasting," 2015.

[10] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "C3D: generic features for video analysis," CoRR, vol. abs/1412.0767, 2014.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489–4497.

[12] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–6.

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," 2015.

[14] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," Computer Ence, 2014.

[15] Michael Mathieu, Camille Couprie, and Yann Lecun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440,* 2015.

[16] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine, "Stochastic variational video prediction," CoRR, vol. abs/1710.11252, 2017.

[17] Ian J. Goodfellow, Jean Pouget-Abadie, et al., "Generative adversarial networks," 2014, cite arxiv:1406.2661.

[18] Channappayya, S., et al., "Rate bounds on ssim index of quantized images," IEEE Transactions on Image Processing, 2008.