Future Frame Prediction for Robot-assisted Surgery

Xiaojie Gao, Yueming Jin, Zixu Zhao, Qi Dou, and Pheng-Ann Heng

Dept. of Computer Science & Engineering, The Chinese University of Hong Kong

IPMI 2021
Rønne, Bornholm

Introduction

Generating detailed future scenes is an advanced stage of surgical video interpretation towards recognizing the current context. In this paper, we study the future frame prediction from robot-assisted surgical videos with dual-arm scenarios. Targeting at the intricate movements of robotic arms, we propose a Ternary Prior Guided Variational Autoencoder (TPG-VAE) model that combines the learned content and motion prior together with the constant class label prior to predict the future frames of dual-arm robots. Results show that our method could capture the meaningful movements of robotic arms and outperforms compared methods for general videos in both quantitative and qualitative evaluations.

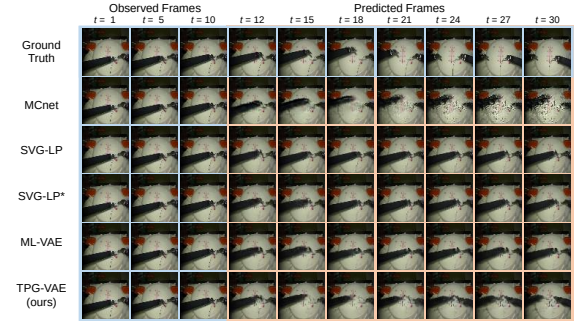

Fig. 3. Qualitative results showing the gesture of G2 (positioning needle) among different models. Compared to other VAE-based methods, our model captures the moving tendency of the left hand while other methods only copy the last ground truth frame. Frames with blue edging indicate the ground truth while the rest are generated by each model.

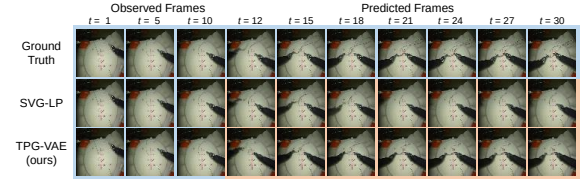

Fig. 4. Qualitative comparison indicating the gesture of G4 (transferring needle from left to right). Our method generates images with high quality and predicts the actual location of the left arm, while the compared method tends to lose the left arm.

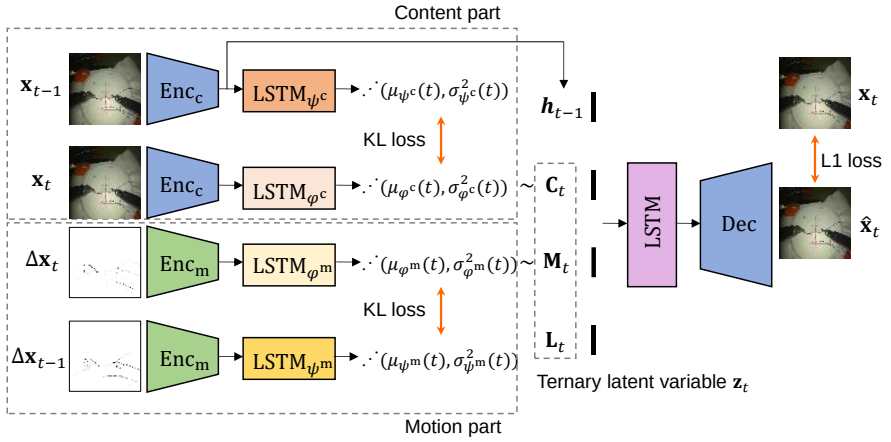

Fig. 1. Illustration of the training process of the proposed model. The posterior latent variables of content and motion at time step t together with the class label prior $\mathbf{L}_t$ try to reconstruct the original frame $\mathbf{x}_t$ conditioning on previous frames. And the prior distributions from content and motion at time step $t-1$ are optimized to fit the posterior distributions at time step t.

Method

- Our core idea is to employ the learned prior from content and motion information together with the constant class prior to constrain the latent space of the model for future frame prediction.
- We firstly design a decomposed video encoding network to capture the content and motion distribution from surgical videos to consider the diverse operation trajectories. The spatial information and temporal dependency among videos are extracted and preserved via CNN and LSTM.
- Next, we devise a ternary latent variable that consists of content, motion, and class information for guiding the decoder to generate the next frame. This is also consistent with the forecasting procedure of humans by referring to various prior.
- The network is trained by maximizing the variational lower bound, and the likelihood term is replaced with L1 reconstruction loss.

Results

- Dataset: Video clips from the suturing task of the JIGSAWS dataset.
- Ablation settings: SVG-LP*: SVG-LP trained using L1 loss and tested without sampling; ML-VAE: our full model without latent variables of content.
- Each model generates future frames conditioning on 10 observed frames.

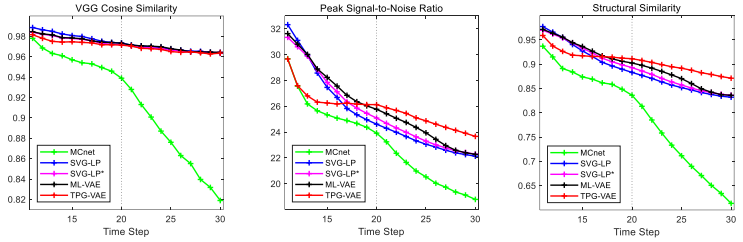

Fig. 2. Quantitative evaluation on the average of the three metrics towards the 100 testing clips. The dotted line indicates the frame number the models are trained to predict up to; further results beyond this line display their generalization ability. For the reported metrics, higher is better.

Table 1. Comparison of predicted results at different time step (mean±std).

Methods	PSNR				SSIM			
	t=15	t=20	t=25	t=30	t=15	t=20	t=25	t=30
MCnet [1]	25.34±2.58	23.92±2.46	20.53±2.06	18.79±1.83	0.874±0.053	0.836±0.058	0.712±0.074	0.614±0.073
SVG-LP [2]	27.47±3.82	24.62±4.21	23.06±4.20	22.14±4.09	0.927±0.054	0.883±0.080	0.852±0.089	0.832±0.088
SVG-LP*	27.85±3.57	25.09±4.13	23.30±4.30	22.30±4.21	0.933±0.046	0.893±0.072	0.857±0.087	0.836±0.088
M-VAE	27.74±3.67	25.14±4.09	23.24±4.30	22.15±4.10	0.932±0.050	0.894±0.072	0.857±0.088	0.834±0.087
CM-VAE	27.44±3.83	25.09±4.07	23.02±4.19	22.16±4.16	0.927±0.056	0.893±0.075	0.853±0.087	0.834±0.088
CL-VAE	28.00±3.73	25.32±4.15	23.49±4.34	22.24±4.28	0.935±0.042	0.897±0.073	0.862±0.087	0.835±0.088
ML-VAE	**28.24±3.51**	25.77±4.02	23.95±4.26	22.28±4.26	**0.936±0.046**	0.903±0.071	0.870±0.084	0.836±0.088
TPG-VAE (Ours)	26.26±3.17	**26.13±3.85**	**24.88±3.68**	**23.67±3.50**	0.917±0.048	**0.911±0.060**	**0.892±0.067**	**0.871±0.071**

Conclusion

We propose a novel method based on VAE for conditional robotic video prediction, which is the first work for dual arm robots. The suggested model employs learned and inherent prior information as guidance to help generate future scenes conditioning on the observed frames. The stochastic VAE based method is adapted as a deterministic approach by directly using the expectation of the distribution without sampling. Our model beats the compared methods on the challenging dual-arm robotic surgical video dataset. Our method is promising to be applied to various surgical scenarios, such as image-guided surgery, surgeon training etc.

Reference

[1] Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. In: ICLR (2017)
[2] Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: ICML (2018)