pHealth 2020
B. Blobel et al. (Eds.)

doi:10.3233/SHTI200621

Surgical Hand Gesture Prediction for the Operating Room

Inna SKARGA-BANDUROVA[a,1], Rostislav SIRIAK[b], Tetiana BILOBORODOVA[b], Fabio CUZZOLIN[a], Vivek Singh BAWA[a], Mohamed Ibrahim MOHAMED[a], R. Dinesh Jackson SAMUEL[a]
[a]*Oxford Brookes University*
[b]*Volodymyr Dahl East Ukrainian National University*

Abstract. Technological advancements in smart assistive technology enable navigating and manipulating various types of computer-aided devices in the operating room through a contactless gesture interface. Understanding surgeon actions is crucial to natural human-robot interaction in operating room since it means a sort of prediction a human behavior so that the robot can foresee the surgeon's intention, early choose appropriate action and reduce waiting time. In this paper, we present a new deep network based on Convolution Long Short-Term Memory (ConvLSTM) for gesture prediction configured to provide natural interaction between the surgeon and assistive robot and improve operating-room efficiency. The experimental results prove the capability of reliably recognizing unfinished gestures on videos. We quantitatively demonstrate the latter ability and the fact that GestureConvLSTM improves the baseline system performance on LSA64 dataset.

Keywords. Hand gesture, prediction, operating room, surgeon, ConvLSTM, GestureConvLSTM

Introduction

Surgical gestures are an important part of non-verbal communication in the operating room. Recent studies [1, 2] show that during operations, the gesture is not just the occasional accompaniment of speech, it is the separate meaningful activity. The hand gestures are widely used by surgeons for representing some objects or actions, e.g. pointing to the particular area of the operating surface, demonstrating how to perform an action (intubate the patient, pull the tissue, etc.). Gestures can be used in coordinated task activities, e.g. manipulating objects or contactless handling the images as well as interaction with different elements of the user interface. Examples include navigating MRI and CT scans in the operating rooms. In these systems, gesture-based interfaces enable surgeons to interact directly with imaging systems, displays, and controllers without breaking the sterility barrier.

Human-robot interaction in the operative room is a special case of communication that required more advanced techniques for implementation. Working with a robot-assistant, we tend to achieve the same level of performance as with an experienced

[1] Inna Skarga-Bandurova, School of Engineering, Computing and Mathematics, Oxford Brookes University, Wheatley Campus, Wheatley, Oxford, OX33 1HX, UK; E-mail: iskarga-bandurova@brookes.ac.uk.

assistant capable to understand the surgeon's intention from visual content. In this scenario, gesture recognition and classification is not enough, since time delays slow down the interaction. Action prediction techniques aimed to recognize the unfinished gestures could provide natural interaction between a surgeon and a robot in an operating room. The gesture prediction is a before-the-fact video understanding task focusing on the future state, it can be used in the following scenarios:

- to estimate surgeon gestures in advance of the actual action to reduce delays in interactive systems
- to anticipate what images the surgeon will need to see next and what instruments will be needed
- to predict the most likely area that the surgeon may want to inspect using the electronic image medical record, and therefore saving browsing time between the images
- to catch and track the surgeon hand trajectory movements to reflect and recognize certain medical functions such as navigation activities, conversational gestures or operations directive.

There are a set of studies dedicated to gesture recognition and prediction [3, 4, 5]. The most substantial results for video data were obtained with deep learning approaches, Generative Adversarial Networks (GAN) [5] and Long Short-Term Memory (LSTM) networks [6]. Particularly, LSTM, as a special deep learning structure, has shown great ability in modeling long-term dependencies in the time dimension of video. However, adapting the gestural recognition system to specific environments like operating rooms requires significant effort. Inspired by the above works, the main purpose of this paper is to construct a novel deep neural network structure for surgeon gesture prediction providing efficient human-computer interface in operating room.

Our goal is to predict the next few frames of a video given the previous ones and to make and assumption about action labels based upon temporally incomplete hand gesture videos. Formally, given an incomplete gesture video $x_{1:t}$ containing t frames, i.e., $x_{1:t} = \{f_1, f_2, \cdots, f_t\}$, the goal is to infer the action label y: $x_{1:t} \rightarrow y$. Here, the incomplete hand gesture video $x_{1:t}$ contains the beginning stage of a complete movement execution $x_{1:T}$, which only contains one single gesture. This video frame prediction task requires at least two types of intelligence. The first one is a visual classifier enabling to interpret each gesture on the frame and the second one is a time sequence modelling to understand the sequence of frames. To solve this task, we propose a sequence generation model using Convolutional LSTM (ConvLSTM) with dropout regularization to reduce overfitting and reduce generalization error in prediction of hand gestures.

1. Methods

The backbone of the proposed solution is the ConvLSTM [6] which extend the fully connected LSTM layer to use convolutional structures both in the transition between inputs and states and between states, so that it is possible to simultaneously study spatial and temporal relationships. We use the open source TensorFlow Keras implementation of the ConvLSTM network presented in [7] as the baseline.

The main difference in our approach is in replacing the three batch normalization layers with dropout regularization layers that enable simulating a large number of networks with different network structure and, in turn, making nodes in the network generally more robust to the inputs [8]. Also, a hand edge detection described in our previous work [9] is used on data preprocessing stage to remove the noise.

The proposed GestureConvLSTM (Figure 1) is mainly composed of three components, ConvLSTM2D class with a 2D filter which computes convolutional operations in both the input and the recurrent transformations, Dropout regularization to reduce overfitting and improve generalization error, and Conv3D class with a 3-D convolution filter to extract a more discriminative feature representation. ConvLSTM2D is a special architecture that combines LSTM gating with two-dimensional packages, it is almost the same as a traditional LSTM [10] layer, except the input and recurrent transformations are two-dimensional convolution transformations. This enables to interpret visual information in the form of two-dimensional images and offers a means for analyzing time sequences which in turn makes it possible to perform predicting / generating video frames.

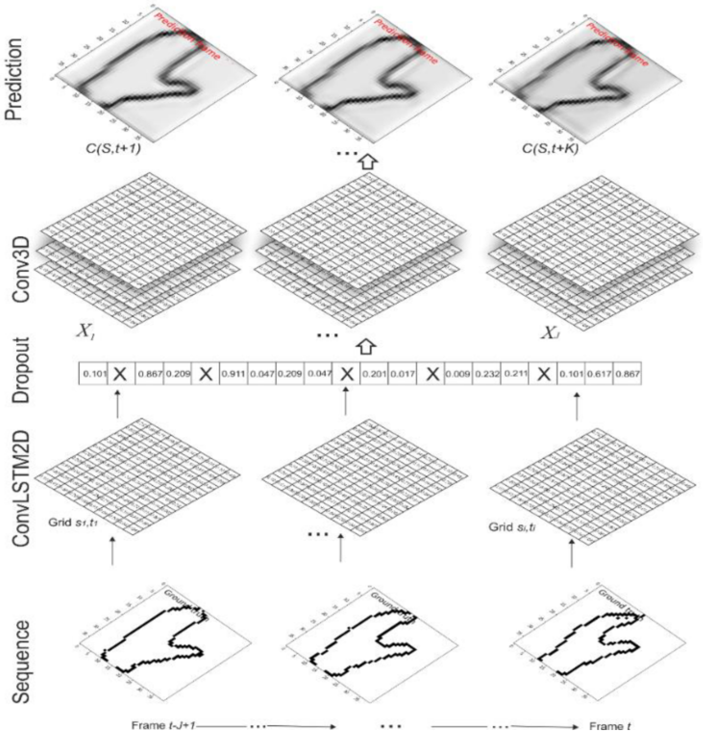


Figure 1. The pipeline of the proposed gesture prediction model.

This model is used to predict the next frames in a sequence of the length of t. The features of frames up to t − 1 are used to predict the features on the next frame.

The GestureConvLSTM architecture consists of four ConvLSTM2D layers, one Conv3D layer, and three dropout regularization layers. It processes each image taken from the camera at a rate of 19 frames per second to produce meaningful features. Additionally, it represents one of the few structures capable of scaling according to the data, i.e. its depth can be easily increased or decreased depending on the scene complexity without suffering from model overfitting during training. The kernel size (3, 3) is maintained throughout all layers to improve feature detection at different scales.

2. Results

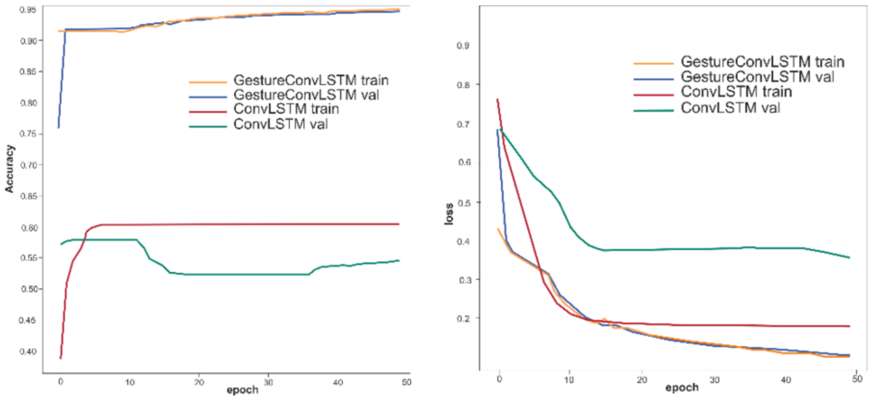
We tested our GestureConvLSTM on a Google Collaboration using GPU graphics [11], the source code of ConvLSTM with batch normalization by Keras [7] used as a control model.

2.1. Datasets

We trained our network on the data from the open-source sign database for the Argentinian Sign Language LSA64 [12]. This dataset includes 3200 videos where 10 non-expert subjects executed five repetitions of 64 different types of signs. Subjects wore black clothes and performed the signs standing or sitting, with a white wall as a background. To simplify the problem of hand segmentation within an image, subjects used fluorescent-colored gloves. These substantially simplify recognizing the position of the hand and remove all issues associated to skin color variations, while fully retaining the difficulty of recognizing the handshape. The resolution of original videos is 1920 × 1080, at 60 frames per second. For our experiment, we used the cropped videos. They are similar to the initial ones but each video was temporarily segmented so that the frames at the beginning or end without hands motion were deleted.

2.2. Training and validation results

The training of the network was conducted on 50 video sequences, each sequence consisting of 30 frames. In order to save resources, the frame size is reduced to 40×40. Thus, all input data were presented as a dimension tensor (50, 30, 40, 40, 1). At the prediction stage, based on the initial ten frames from the sequence that did not participate in the training, ten more subsequent frames are generated, which should predict the further movement of the hand. The rate of generation of the predicted sequence is 19 frames per second. The network performance was evaluated using accuracy, i.e. the percentage of correctly labelled frames and binary cross-entropy / log loss of each predicted sequences. We also trained ConvLSTM with the original annotations of LSA64, in order to compare our GestureConvLSTM model to the baseline [6]. Figure 2 shows recognition accuracies of individual gestures and binary cross-entropy / log loss of each predicted sequences on training and validation datasets for regular ConvLSTM and GestureConvLSTM.



(a) **(b)**

Figure 2. Results of training and validation: (a) accuracy and (b) binary cross-entropy / log loss for one gesture from LSA64, ID1: Opaque.

As can be seen, the proposed GestureConvLSTM converged much faster and obtained an accuracy of close to 94% on the validation set, whereas for the same data, the regular ConvLSTM model plateaued on 60% around the fifth epoch. The validation loss is significantly lower than that obtained using the regular model. The proposed GestureConvLSTM ensures a fast yet stable convergence.

2.3. Evaluation Metrics and Prediction Performance

The prediction performance of the GestureConvLSTM and ConvLSTM were evaluated using qualitative and quantitative approaches. The frames are paired image-to-image and comparison results are shown in Figure 3 and Table 1.

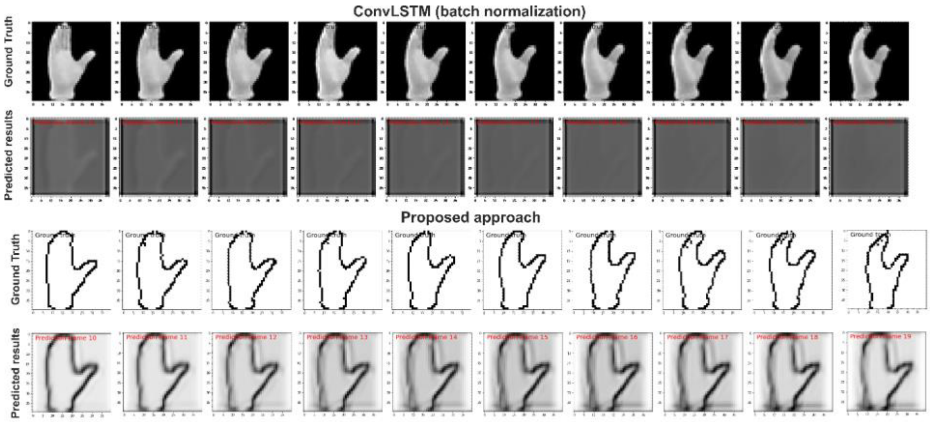


Figure 3. Visualization of the 10-frame ahead ground truth with predicted gesture obtained from ConvLSTM with batch normalization and from proposed ConvLSTM with dropout and contouring.

Results are shown for a gesture ID1: Opaque (performed with the right hand).

Table 1. Gesture prediction results on the LSA64 dataset. Euclidean distance, Manhattan Distances (MD), and normalized cross-correlation are used as evaluation metrics.

Metrics	**Euclidean Distance**		**Manhattan Distance**		**Normalized Cross Correlation**	
	ConvLSTM	GestureConvLSTM	ConvLSTM	GestureConvLSTM	ConvLSTM	GestureConvLSTM
Frame 10	0.00061	0.00071	0.13987	0.11838	0.09638	0.61510
Frame 11	0.00061	0.00073	0.13877	0.11946	0.06622	0.51900
Frame 12	0.00059	0.00076	0.13886	0.11958	0.03272	0.42948
Frame 13	0.00059	0.00081	0.14032	0.12236	0.01243	0.35369
Frame 14	0.00060	0.00080	0.14173	0.11803	-0.00358	0.28873
Frame 15	0.00060	0.00080	0.14172	0.11503	-0.01721	0.23654
Frame 16	0.00060	0.00080	0.14348	0.11152	-0.03392	0.22801
Frame 17	0.00061	0.00081	0.14418	0.10743	-0.03822	0.17827
Frame 18	0.00061	0.00081	0.14448	0.10598	-0.04318	0.14165
Frame 19	0.00060	0.00081	0.14305	0.10363	-0.06131	0.14899
Average	0.00060	0.00078	0.14165	0.11414	0.00103	0.313946

3. Discussion and Conclusion

Visually, the GestureConvLSTM produces more photo-realistic results than a regular network. With a qualitative comparison of the predictions, it is noticeable that more accurate movement of the hand was obtained using the GestureConvLSTM what matches with the results obtained using normalized cross-correlation while Euclidean distance for our approach is largely unfit for the purpose. It can be explained the high dimensionality. According to [13] for these data, the MD metric is the more preferable over the Euclidean distance. The average normalized cross-correlation of 10 predicted successive frames for GestureConvLSTM was 0.313946 versus 0.00103 for ConvLSTM that indicates a better similarity between the predicted and the real frames for GestureConvLSTM.

This is research is the first step towards understanding the surgeon intends. The results allow for the conclusion that GestureConvLSTM can be used for estimating gestures in advance of their actual action reducing delays in interactive systems. This innovation, together with other novel approaches such as voice recognition, eye-tracking, tracking a surgical instrument inside the patient's body, other assisted technologies might significantly improve the efficiency of an operating room and help to reduce the potential infection and the length of surgeries. The future improvements are directed on understanding the context and interaction during surgical operation in which gestures are made and discriminating between intended gestures versus unintended gestures.

Acknowledgement

This paper is supported by European Union Horizon 2020 research and innovation program under grant agreement No 779813, project SARAS (Smart Autonomous Robotic Assistant Surgeon).

References

[1] Bezemer J. Gesture beyond conversation Handbook of Multimodal Analysis, ed. C. Jewitt (2010).

[2] López-Casado C, Bauzano E, Rivas-Blanco I, Pérez-Del-Pulgar CJ, Muñoz VF. A Gesture Recognition Algorithm for Hand-Assisted Laparoscopic Surgery. Sensors (Basel). 2019 Nov 26;19(23):5182. doi: 10.3390/s19235182. PMID: 31779237; PMCID: PMC6929113.

[3] Lee AR, Cho Y, Jin S, Kim N. Enhancement of surgical hand gesture recognition using a capsule network for a contactless interface in the operating room. Computer Methods and Programs in Biomedicine. 2020 Jul;190:105385. DOI: 10.1016/j.cmpb.2020.105385.

[4] van Amsterdam B., Clarkson M.J., Stoyanov D. Multi-task reccurent neural network for surgical gesture recognition and progress prediction. In 2020 IEEE International Conference on Robotics and Automation (ICRA). https://arxiv.org/pdf/2003.04772.pdf

[5] Tang H., Wang W., Xu D., Yan Y., Sebe N. GestureGAN for Hand Gesture-to-Gesture Translation in the Wild In 2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA. https://doi.org/10.1145/3240508.3240704

[6] Xingjian, S. H. I., et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems. 2015.

[7] Keras https://keras.io/examples/conv_lstm/

[8] Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I. and Salakhutdinov R. R. Improving neural networks by preventing co-adaptation of feature detectors arXiv:1207.0580v1 [cs.NE] 3 Jul 2012

[9] Siriak R., Skarga-Bandurova I., Boltov Y. Deep Convolutional Network with Long Short-Term Memory Layers for Dynamic Gesture Recognition 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2019. – IEEE, 2019; 1: 158-162.

[10] LeCun, Y. and Bengio, Y., 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), p.1995.

[11] Google Collaboration using GPU graphics https://colab.research.google.com/notebooks/intro.ipynb]

[12] Ronchetti F., Quiroga F., Estrebou C., Lanzarini L., Rosete A. LSA64: An Argentinian Sign Language dataset, XX II Congreso Argentino de Ciencias de la Computación (CACIC), 2016.

[13] Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In International conference on database theory, pp. 420-434. Springer, Berlin, Heidelberg, 2001.