Title:

VGG-3DConvNet for surgical action recognition

Authors:

Razeen Hussain[1], Raabid Hussain[2], Youssef Skandarani[2], Kibrom Berihu Girum[2]

Affiliations:

[1] DIBRIS, University of Genoa, Genoa, Italy

[2] ImViA Laboratory, University of Burgundy, Dijon, France

Team name (if any):

Parakeet

Do you want to be part of the challenge summary publication by making this report public?

Yes

Introduction

Automatic surgical action recognition is a promising direction for creating next generation intelligent surgical devices and support systems. Such systems can potentially enhance the surgical workflow and efficiency and result in lower costs and improved patient care. However, it is not easy to acquire surgical data due to much sensitivity related to data privacy. Therefore context-aware algorithms need to be developed to overcome this challenge. This work focuses on proposing an automatic visual domain adaptation algorithm for action recognition in the operating room. The challenge focuses on training action recognition models on virtual data from simulated surgical tasks and applying them on real world surgical applications.

The training dataset consisted of videos from two domains: 256 virtual reality (VR) and 26 clinical-like videos of porcine models captured from a da Vinci robotic system. Three processes were being performed in the videos: dissection, knot-tying, and needle driving. The testing dataset consisted of 16 video clips from the porcine model. The challenge was divided into two main tasks. Task 1 comprised of training a convolutional neural network (CNN) model using VR and porcine videos whereas for task 2, the model had to be trained only on VR videos. Both task's models were tested on porcine videos.

Since, the data in the training and testing datasets of the challenge was significantly different, we proposed to incorporate both spatial and temporal information in our approach. We hypothesize that both these types of information can be useful. Our proposed methodology first extracts pre-trained spatial VGG16 features [1] from the video frames and then extracts temporal information through a 3DConvNet [2] based on continuous video frames. The proposed methodology is explained in the following section in detail.

Methodology/Results

Since the fps of the VR and porcine videos was significantly different, the training videos were downsampled to 5 fps. This frame rate was chosen as the videos consisted of moderate movements which do not require a high refresh rate for determining movements. For memory limitations, the frames were also desampled to a lower resolution of 224x224x3 pixels. The video frames were input to a 2D-3D CNN depicted in Figure 1. The proposed architecture takes sequence of 8 frames as input (8x224x224x3) and outputs the softmax classification scores. First 2D VGG16 features (7x7x512) were extracted for each video frame individually. The VGG network was pretrained using the imagenet dataset [3]. The VGG16 features for the sequence were passed through a 3DConvNet to integrate spatial and temporal information. The 3DConvNet consisted of 3x3 convolutional layers followed by a fully connected layer and was trained from scratch. All convolutional layers had ReLU activations followed by batch normalization. Categorical-crossentropy was used as the loss function. A 40% dropout was used before the final fully connected softmax layer. The filter sizes are shown in figure 1.

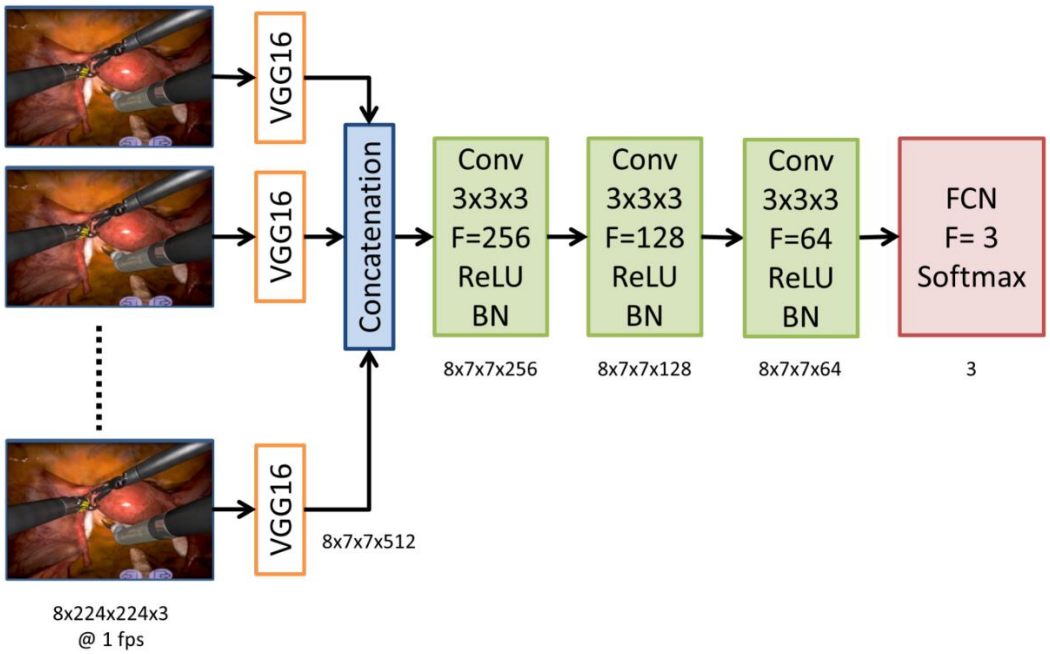

Figure 1: Proposed spatio-temporal architecture. VGG16: 2D VGG16 pretrained on imagenet dataset, Conv: 3D convolutional layer, BN: batch normalization, FCN: fully connected layer, F: number of features.

The architecture was implemented on a computer with dedicated GPU (NVIDIA GeForce GTX 1080, 8 GB RAM processor) using Keras and Tensorflow libraries. For task 2, only a training dataset (without validation) was used containing all the VR video frames. Whereas for task 1, some porcine model videos were included in the training dataset along with VR videos with a training-testing split ratio of 60:40. The training was performed for 500 epochs with a batch size of 32 using Adam optimizer with a learning rate of 0.0001.

During training, a data augmentation strategy was adopted in which the video stream was divided into fixed size segments. The input frames for the proposed VGG-3DConvNet were randomly selected from within each segment. Thus, the input to the network was at 1 fps with data augmentation performed on 5 fps data stream. The validation dice scores for the individual tasks were: 88% for task 1 and 69% for task 2.

Conclusion/Discussion

We proposed a VGG and 3DConvNet based automatic architecture for surgical action recognition. The proposed architecture combines spatial information, by first extracting individual 2D features from each video frame, with temporal information by passing the spatial features through a 3D CNN. The architecture yielded good performance on dissection and needle driving cases. The worst performance was in the case of knot-tying as spatial information was also integrated into the network and most of the training VR videos were very different from the porcine model videos in the testing dataset; whereas, the other two surgical tasks were predicted correctly most of the times. In future, we propose to use the virtual videos to automatically generate simulated data resembling the porcine videos using generative adversarial networks to increase the augmentation of the data [4]. Another direction could be to integrate prior shape knowledge of the instruments to better identify temporal movements [5].

References

[1] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[2] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

[3] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[4] Skandarani, Y., Painchaud, N., Jodoin, P.M. and Lalande, A., 2020. On the effectiveness of GAN generated cardiac MRIs for segmentation. arXiv preprint arXiv:2005.09026.

[5] Girum, K.B., Lalande, A., Hussain, R. and Créhange, G., 2020. A deep learning method for real-time intraoperative US image segmentation in prostate brachytherapy. International Journal of Computer Assisted Radiology and Surgery, 15(9), pp.1467-1476.