Felix Yu - Group 2

4/26/2018

EN.601.656, Spring 2018

Critical Review of

Surgical Phase Recognition: from Instrumented ORs to Hospitals Around the World[1]

Project Goals

The ability to analyze surgical videos has many uses, ranging from optimizing the efficiency of the surgical setup, analyze the workflow of the procedure, and provide feedback for the surgeons. Because of this, a recently developed area of research involves bringing video processing techniques from other fields into the domain of surgical videos. Specifically, various neural network approaches are being explored in order to perform surgical phase recognition, a problem that consists of segmenting a video of a surgery into it's various phases, that are performed during the course of the specific surgery. These segments will be referred to as activity clips. From there, the goal of group 2's Query by Video project is to encode activity clips into vectors that can be used to determine how similar a clip in the database is to the query clip.

Selected Paper

The selected paper by Lea, C., et. al was selected for multiple reasons. Although the problem the paper tries to solve, which is the segmentation of the surgery, is different than our query by video problem, the former problem must be solved to a reasonable degree in order for the latter problem to be feasible. This relationship is described above. Furthermore, not only is the data domain almost identical, but the pipeline and methods used in Lea's paper provide inspiration for our own project. Specifically, Lea also explores the use of encoding images first into spatial features, which then are fed into a temporal convolutional network in order to capture temporal features as well. Furthermore, Lea also includes tool annotations into his model, which is related to our maximum deliverable.

Key Results

There are three main contributions made by the paper.

1. The efficacy of using a spatio-temporal CNN to extract features from the video is analyzed.
2. Various phase classifiers are compared to each other, with each one given the outputs of the ST-CNN.
3. A new and messier dataset, EndoTube, is introduced.

Background Information

Although the findings of the paper can be applied to any surgery given an availability of dataset, the videos analyzed by the paper's pipeline are cholecystetomy videos. There are two datasets that are looked into, the first is the EndoVis dataset[1], while the other is the EndoTube dataset which was generated by the authors of the paper themselves in order to explore how models performed under data that is poorly standardized. All datasets come with videos of the surgeries, ground truth phase annotations for each frame of the surgery, as well as ground truth tool

annotations for each frame. The goal of the paper is to use this data in order to segment the video into it's respective phase clips. A visualization for the data is given below. The eight images are frames taken from various activities during the cholecystetomy video, and the colored bar indicates ground truth phase labels for each frame of the data, for the duration of the video.
The colored bar is what Lea's model tries to predict.

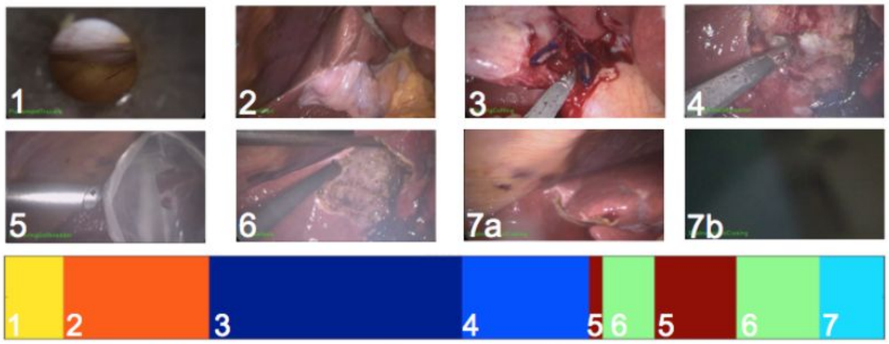

Fig. 1: Example images and sequence labeling from the EndoVis dataset. Phases: (1) Place trocars (2) Prepare Calots triangle (3) Clip/cut cystic artery and duct (4) Dissect Gallbladder (5) Retrieve Gallbladder (6) Hemostasis (7a/7b) Drainage/closure/finish.

Data Collection/Methods

The EndoTube dataset was created by finding whole cholecystectomy procedures on Youtube, and then hand-labeled by those involved with the project to have the same phases as those used in EndoVis. The dataset has 25 videos, performed across 19 hospitals and 9 countries. Various clips that were not a part of the surgery themself were kept, but given no label.
There are three stages to Lea's model designed to solve the segmentation problem. The first stage encodes each frame of the RGB image into spatial feature vectors through a CNN. The second stage takes these outputs as inputs and uses another shallow CNN to capture short range (60 second) temporal dependencies as well. Finally, the third stage takes the feature vector outputted by the temporal CNN to run phase classification on each frame of the video, therefore segmenting the image. Each portion is described more in depth below.

The Spatial CNN

The spatial CNN is straightforward. This neural network, with an architecture based off of VGG[3], takes in a single image and then outputs a feature vector which is trained to predict either tool annotation (if available), or phase label (if tool annotation is not available).

The Temporal CNN

Again, the temporal CNN is relatively straightforward. The encoding portion of the network is only one layer thick, and consists of 32 filters that convolute windows that span 60 seconds. In other words, a sliding window summarizes the spatial features corresponding to frames that span 60 seconds, and outputs each of these summaries (there are still as many summaries as there are frames in the image, since this is a sliding window). If there are tool annotations available, the feature vector for that is then concatenated onto these summaries at each timestep, and these become the input to the phase classifier.

The Phase Classifier

Three classifiers are tested.

1. (LM) The Linear Model uses a multi-class logistic regression (softmax) to directly map each spatio-temporal encoding onto a feature class. This assumes each time step's class is independent of the class at any other time-step.
2. (SMM) The Semi-Markov Model is adapted into a segmental model developed by Lea *et. al.* in [9] in order to take into account how different phases transition into one another, as well as how features in consecutive time-steps relate to one another.
3. (DTW) Dynamic Time Warping is used to make the feature vectors that exist across time-steps time invariant, and then a nearest neighbors approach is used to classify which phase clumps of signals belong to.

Below are diagrams that represent each portion of the pipeline detailed above.

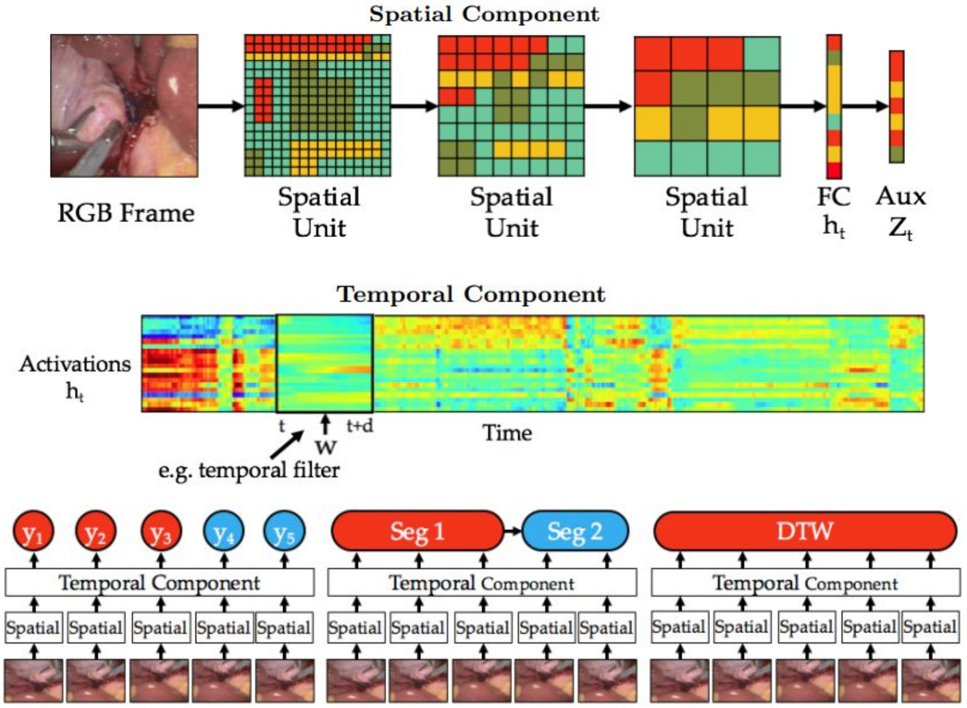

Fig. 3: The full models with the Spatiotemporal CNNs and classifiers. (left) Linear Model (middle) Segmental Model (right) Time-invariant Model

Results

The paper uses two metrics in order to analyze accuracy of the pipeline.

1. Per frame accuracy of classification

2. Accuracy of the boundaries within a specific threshold. If a boundary is marked at a specific time that is within a certain number of frames away from a true boundary, the boundary is marked as accurate.

These accuracies are then analyzed for all of the following combinations:

- Spatio-temporal component: Using only the spatial CNN, or using both spatial and temporal
- Phase Classifier: Using LM, SMM, or DTW.
- Data: Using only videos, using only tool annotations, and using both.

Furthermore, two existing published models were also used as comparison.[4] [5]

The following points summarize the results:

1. Using a spatio-temporal neural network to encode features rather than a purely spatial encoder leads to higher accuracy in phase segmentation.
2. Using Dynamic Time Warping leads to higher accuracy as well.
3. Purely tool information creates predictions with higher accuracy than purely video. However, having both increases accuracy the most.
4. The model that uses ST-CNN with DTW outperforms the two existing published models regardless of whether only video, only tools, or all data is used.
5. The models all have lowered performance on the messier EndoTube dataset, and are currently at a level that is not satisfactory.
6. Using ST-CNN with DTW, the boundaries predicted are relatively accurate (85% accuracy if the threshold is within 30 frames, 100% accuracy for 180 frames).

Assessment

The paper does a great job describing all methods used, to the point where if one were to create the models from scratch, it would not be difficult to reproduce the results shown from the paper itself. Furthermore, the results show the effectiveness of having frame-by-frame tool annotations as well as the standard video data. Because of the results of this paper, our group has decided to try and incorporate tool annotations into our model as well. Some additional perks of the model described by the paper, aside from the advantages gained by the modularity of each portion of the model and the accuracy, is the quickness in which the model trains. By using the EndoTube dataset as well, the paper shows the current limitations to its own model, as well as other existing models.

There are portions of the paper tphat seemed weaker, however. The description of the EndoTube dataset was rather weak, and although the paper talks about the increased variability in the data, there are no diagrams that give examples on how much more variable the EndoTube data was in comparison to EndoVis. Furthermore, although the paper does run comparisons between the model presented and existing models, this sample size is relatively low. This makes it difficult to judge whether these new results are truly state of the art. In order to draw a more convincing argument, the authors would have to go more in depth with the comparisons between their proposed model and all existing methods of solving this problem.

References:

[1] Lea, C., et. al. “Surgical Phase Recognition: from Instrumented ORs to Hospitals Around the World.” *Paper presented at M2CAI workshop,* (2016).

[2] TUM EndoVis. http://endovissub-workflow.grand-challenge.org/ (2015).

[3] Simonyan, K., zisserman, A.: “Very deep convolutional netowrks for large-scale image recognition.” *Paper presented at ICLR,* (2015).

[4] Dergachyava, O. et. al. “Automatic data-driven real-time segmentation and recognition of surgical workflow”. *Paper presented at IJCARS,* (2016).

[5] Twinanda, A.P. et. al. “Endonet: a deep architecture for recognition tasks on laparoscopic videos.”. abs/1602.03012 (2016).