

Automatic Detection of Surgical Phase Boundaries in Laparoscopic Cholecystectomy Videos

Babak Namazi PhD^{1,2}, Ganesh Sankaranarayanan PhD², Venkat Devarajan PhD¹, James W. Fleshman²

¹ University of Texas at Arlington, Arlington, TX, USA , ² Baylor University Medical Center, Dallas, TX, USA

INTRODUCTION

- The recorded videos in laparoscopic surgeries offer a great opportunity for education and training future surgeons and can be used for the assessment of surgical skills.
- In order to take advantage of laparoscopic videos, they need to be labeled, organized, and stored.
- Manual labeling of these videos, which might be hours long, can be very time-consuming, requires expert supervision, and is prone to human error or subjective evaluation.
- Automatically extracting the critical parts of videos can be extremely beneficial in designing an information retrieval system.
- An automated analysis system will allow the long video recordings to be quickly cued.
- As a part of such an analysis system, an automated phase detection subsystem is needed.
- Directly detecting the boundaries of each phase is an alternative solution for video segmentation.
- Our goal is to design an automated system based on deep learning techniques to identify the beginning and ending of each phase in a laparoscopic cholecystectomy video.
- In an information retrieval system, a specific phase or task can be requested upon a query, to be further analyzed manually or with the aid of an automated system.

BACKGROUND

- Deep learning is a class of machine learning algorithms that consists of a cascade of multiple non-linear layers for extracting specific patterns from data.
- In supervised learning, a deep learning model is trained with a large set of manually labeled input data in order to make predictions on new and unseen data.
- Convolutional neural networks (CNNs) and Recurrent neural networks (RNNs) are the main components of most deep learning models.
- CNN: a deep feed-forward artificial neural network that is trained to learn high-level spatial features in images and videos
- RNN: a recurrent model that is trained to learn the dependencies in a sequence of inputs

DATASET

- 80 videos from cholecystectomy surgery (cholec80 dataset)
- All videos have been annotated manually
- 40 videos were used for training
- 40 videos were used for validation
- The frames are labeled every 25 frames
- Table 1 shows the duration of all 7 phases

Table 1: The duration of phases in cholec80

	Phase	duration(sec)
P0	Preparation	125 ± 95
P1	Calot Triangle Dissection	954 ± 538
P2	Clipping and Cutting	168 ± 162
P3	Gallbladder Dissection	857 ± 551
P4	Gallbladder Packing	98 ± 53
P5	Gallbladder Retraction	83 ± 56
P6	Cleaning and Coagulation	178 ± 166

METHODS

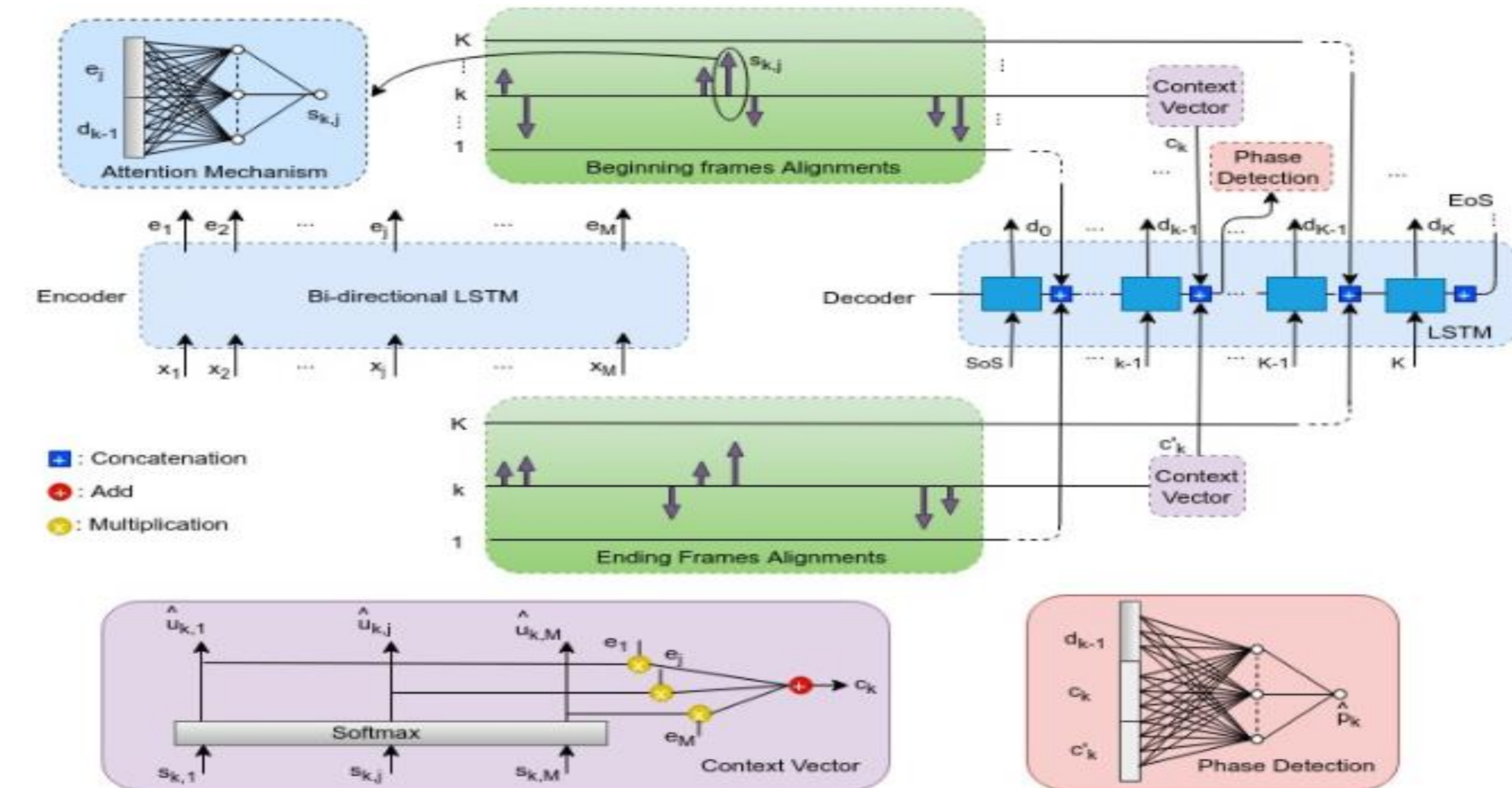


Figure 1: The Block diagram of the attention-based phase boundary detection

- Model:
 - In order to find phase boundaries in a surgical video, we associate a tuple of integers to each phase, which is the indices (frame numbers) of the first (the beginning) and the last (the ending) frames.
 - We proposed a novel method based on the sequence-to-sequence (seq2seq) model and attention mechanism.
 - Seq2seq models aim at mapping variable-length input sequences to another variable-length target sequences. They consist of an encoder component and a decoder component and is also called an encoder-decoder architecture.
 - The role of the encoder in a basic seq2seq model is to encode the input into a fixed-size vector called context vector. The context vector represents the whole sequence and is used in the decoder as the initial state to generate the output sequence.
 - Inspired by human's intuition, attention is defined by the importance of each element of input in making a particular decision at the output.
 - With the help of attention, the alignment of the elements in the input and output sequences is used to determine the dependencies between the sequences.
 - In a seq2seq model with attention, the fixed-size context vector is replaced by a dynamic context vector, which is determined by assigning different weights to all the elements of the input using the attention mechanism.
 - Two attention models were used for the beginning and ending frames and their outputs were concatenated to form the input at each decoder state.
 - The inputs to the attention-based boundary detector model were the visual features that were extracted using a CNN
- Training:
 - The training of attention model was performed in an end-to-end fashion.
 - The loss function for training was the weighted sum of softmax cross-entropy for all the outputs.

RESULTS

- To evaluate the performance of the proposed model, we used the Mean Absolute Error (MAE) in seconds or frames (one frame per second).
- The results for each phase in the cholec80 dataset are shown in Table 2.

Table 2. Mean absolute error in seconds for each phase in cholec80 dataset

	P0	P1	P2	P3	P4	P5	P6	Mean
Beginning frame	0.0	27.67	64.65	108.22	46.57	48.07	71.87	52.53
Ending frame	33.25	41.6	65.87	38.82	58.05	0.0	81.22	45.54

- We also calculated the percentage of the predicted boundaries within different ranges, i.e., the accuracy of the model when different ranges for error are acceptable.
- The results are shown for the beginning and the ending frame in Figures 2 and 3, respectively.

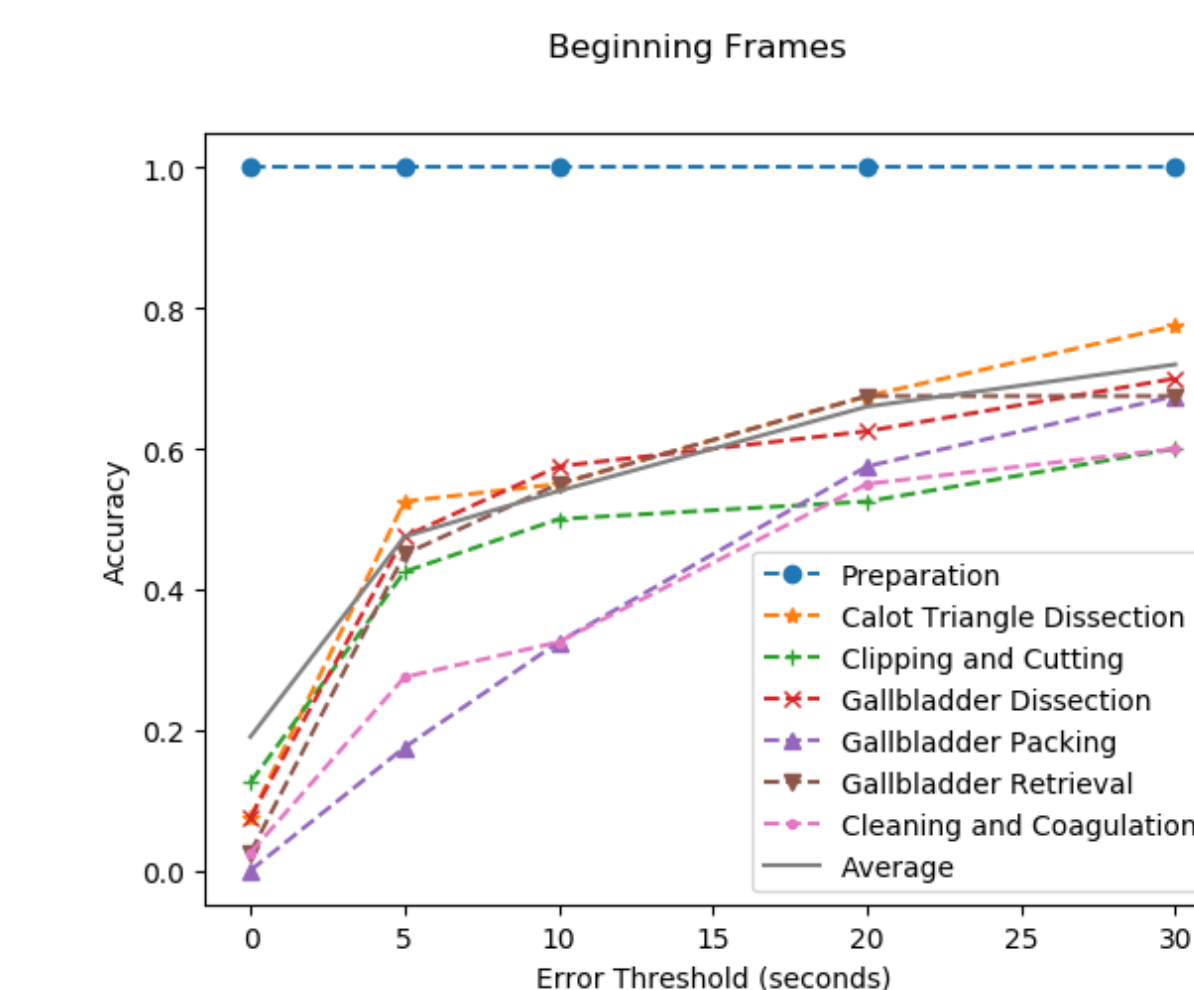


Figure 2: The accuracy of the beginning frames detection for different ranges of error in seconds

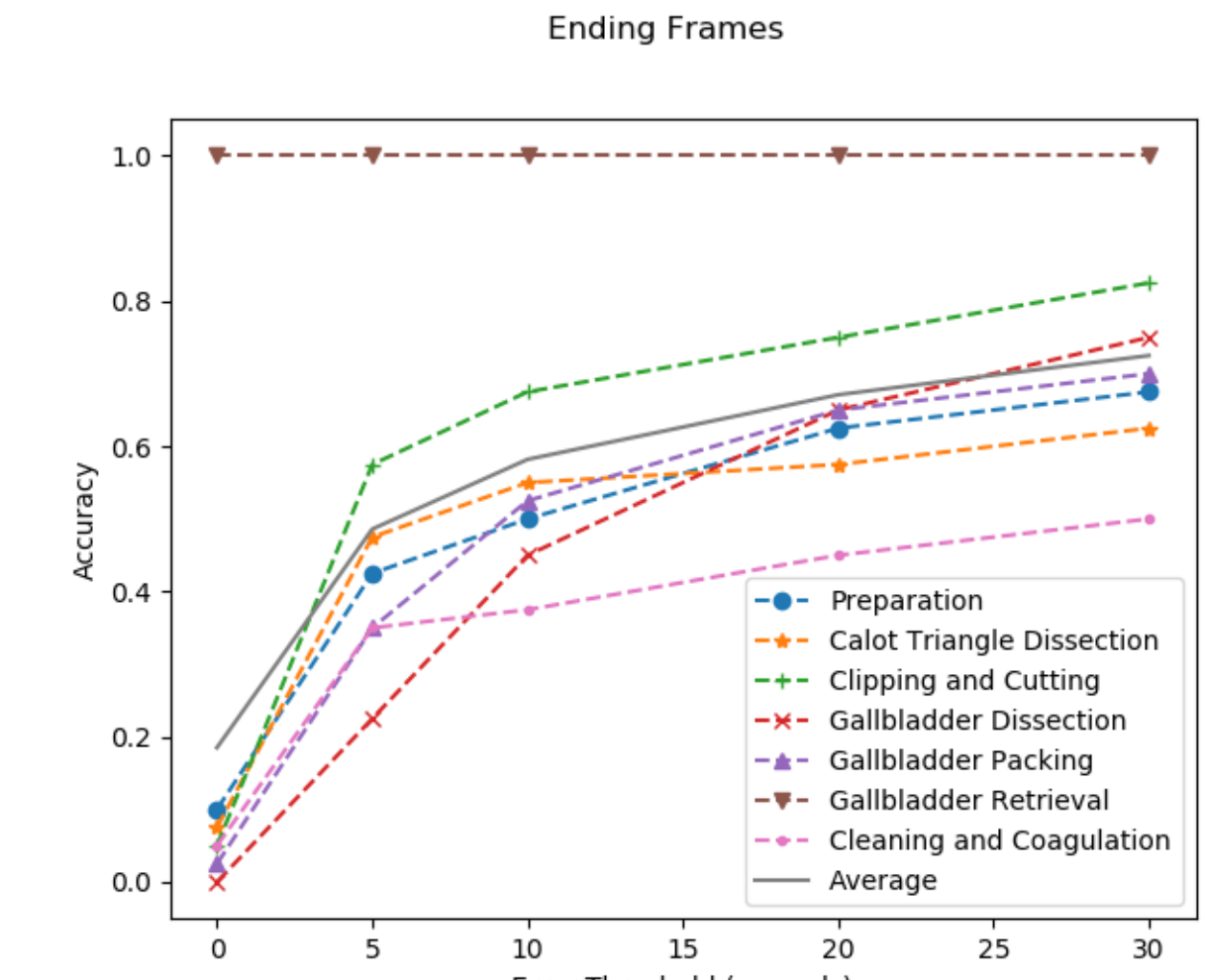


Figure 3: The accuracy of the ending frames detection for different ranges of error in seconds

- It can be seen that more than 48% of the predicted boundary pointers have less than 5 seconds error with respect to the ground truth, whereas the exact match accuracy is less than 19%. This is likely because the transition from one phase to another might take a few seconds.
- The output of the phase detection block had 100% accuracy, which means the model predicted the ordering of the phases correctly.

CONCLUSIONS AND FUTURE DIRECTIONS

- The proposed deep learning model successfully segments a surgical videos into phases and does not require frame-by-frame classification.
- The novel architecture can potentially be adopted for other surgical procedures as well as various video analysis tasks such activity localization, key-frame extraction, summarization, etc.

ACKNOWLEDGMENT

- This work was supported by a Joseph Seeger Surgical Foundation award from the Baylor University Medical Center at Dallas.