Surgical Phase Recognition and Segmentation[1]

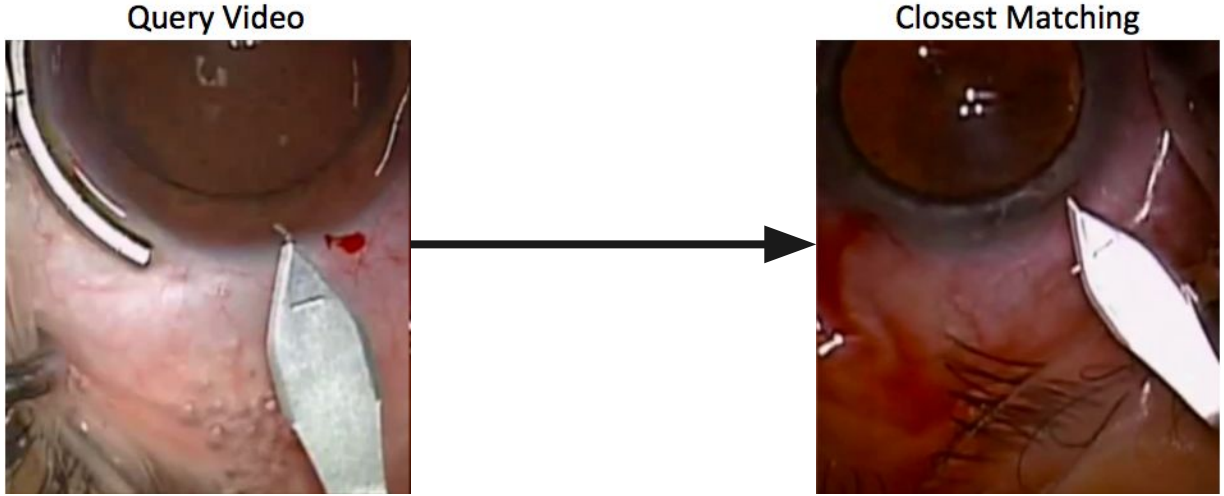
Presented by Felix Yu: Group 2 (Query by Video for Surgical Activities)
Team Members: Felix Yu, Gianluca Croso
Mentors: Tae-Soo Kim, Swaroop Vedula, Gregory Hager, Haider Ali

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

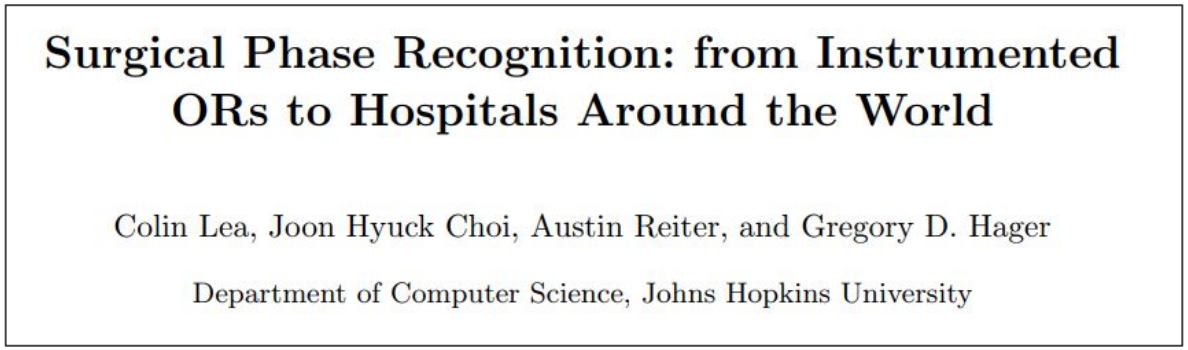
Goals for the Query by Video Project

Design machine learning pipeline to

- Query by video for similar activity from database
- Incorporate tool label information to enhance query results



The Chosen Paper



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

Colin Lea, Joon Hyuck Choi, Austin Reiter, and Gregory D. Hager

Department of Computer Science, Johns Hopkins University

Relevance to the project:

- This problem is one of the parts of the overarching project
- Discusses methods of capturing spatio-temporal information
- Includes methods and analysis of having additional tool-label data

Overarching Project:

- Multiple other portions of the project:
  - Segmentation of whole surgery video into activity clips.
  - **Finding activity clips in database that are similar to the query clip.**
  - Encoding surgeon commentary of database videos into features.
  - Constructing new feedback for query video using the features of similar database videos.

Summary of Paper's Contributions

- Analyze the use of Neural Networks to encode spatio-temporal information that represent surgical phases
- Show results for three classifiers in capturing long term temporal information
- Introduce a new and harder dataset, EndoTube, and analyze results on it.

Background Information

A surgery contains multiple stages, and each one can be analyzed separately.

Question: How can we properly segment a video such that various phases of the surgery are separated from each other?

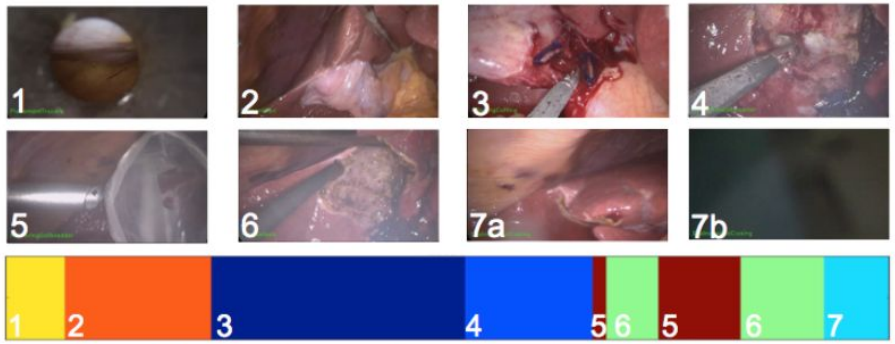


Fig. 1: Example images and sequence labeling from the EndoVis dataset. Phases: (1) Place trocars (2) Prepare Calots triangle (3) Clip/cut cystic artery and duct (4) Dissect Gallbladder (5) Retrieve Gallbladder (6) Hemostasis (7a/7b) Drainage/closure/finish.

Data Information

Data: Two datasets, EndoVis and EndoTube

- Cholecystetomy videos

RGB videos - T frames

Auxiliary Signals (Optional) - One vector per frame

0 0 0 1 0 0 0 0 0 0 1

Phase Label - One label per frame

Data Gathering for EndoTube

25 Whole Cholecystetomy Videos

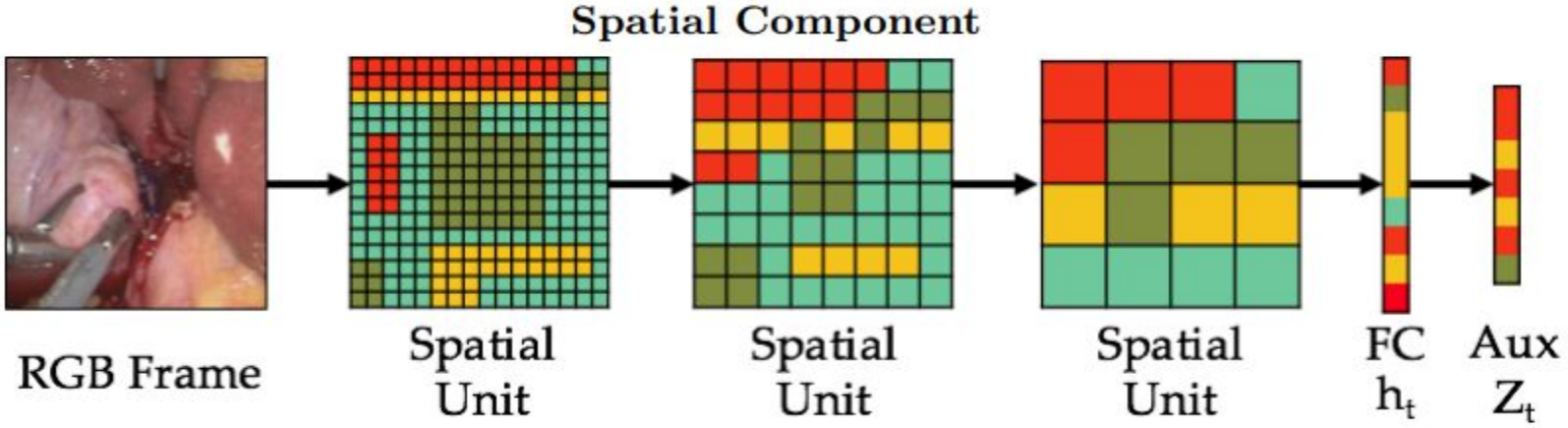
- Taken from YouTube
- Across 19 hospitals in 9 countries
- Phase labels and tool annotations were done by hand
- Annotations were done to match those in EndoVis

The Paper's Approach

Three Parts:

- Spatial Encoding
- Temporal Encoding
- Long-term temporal processing

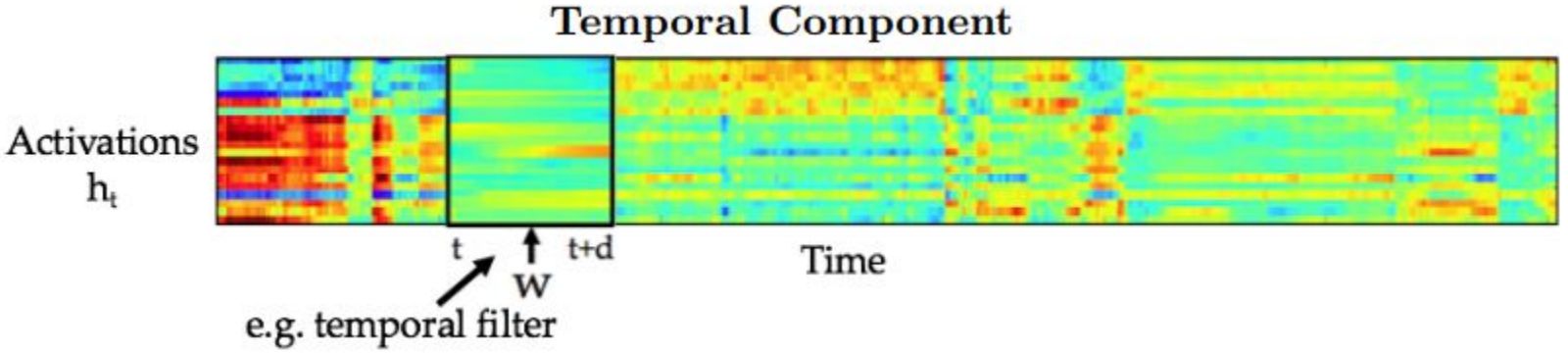
The Paper's Approach: Spatial Encoding



1. 3x3 Convolutional Filters with ReLU activation
2. 3x3 Max Pooling
3. Repeat 1-2 two more times.
4. Fully connected layer into log reg prediction of auxiliary signal (or frame class).

Output of FC layer used as input for next portion.

The Paper's Approach: Temporal Encoding



1. Set of temporal convolutional filters with ReLU Activation (32 filters, each across 60 seconds).
2. Fully Connected Layer into softmax to predict class of each portion of time.

Output of FC Layer used as input to the next portion. The auxiliary signal information is tacked on as well.

The Paper's Approach: Phase Classifier

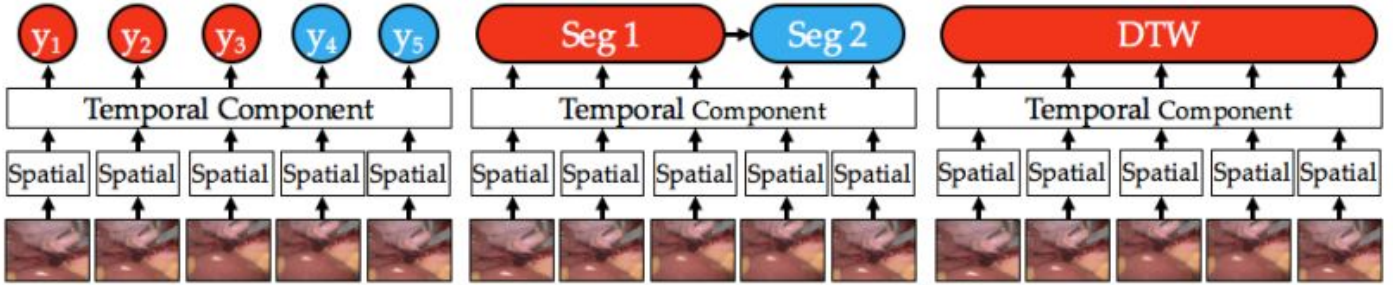


Fig. 3: The full models with the Spatiotemporal CNNs and classifiers. (left) Linear Model (middle) Segmental Model (right) Time-invariant Model

Three Different Models:

1. (LM) Direct output of the spatio-temporal model.
2. (SMM) Semi-Markov Model to better capture interactions between frames and segments.
3. (DTW) Time-invariant Model based off of nearest neighbors to scale signals better.

Results

	Spatial CNN			ST-CNN				
Data source(s)	LM	SMM	DTW	LM	SMM	DTW	[3]	[15]
Video	57.6	78.8	81.2	69.0	77.8	**84.6**	68.1	79.7*
Tools	58.5	76.5	85.7	56.4	78.3	**91.2**	78.9	73.0
Video + Tools	73.7	87.3	92.3	81.8	88.5	**92.8**	88.9	-

EndoVis

	Spatial CNN			ST-CNN		
Data source	LM	SMM	DTW	LM	SMM	DTW
Video	47.9	36.0	63.7	56.3	60.1	62.4

EndoTube

Table 1: Results from (top) EndoVis and (bottom) EndoTube. *[15] achieves 86.0% on EndoVis when pre-training their CNN on a larger dataset and with tool information.

Data source(s)	≤ 30	≤ 60	≤ 90	≤ 120	≤ 150	≤ 180
Video	66.2	76.1	82.5	88.8	93.6	93.6
Tools	**90.4**	**90.4**	92.0	93.6	93.6	95.2
Video+Tools	85.2	**90.4**	**92.0**	**95.2**	**98.4**	**100.0**

Table 2: The percentage of predicted label boundaries within the specified distance (in seconds) to the true boundaries on EndoVis using the DTW model.

Assessment of Paper

Pros:

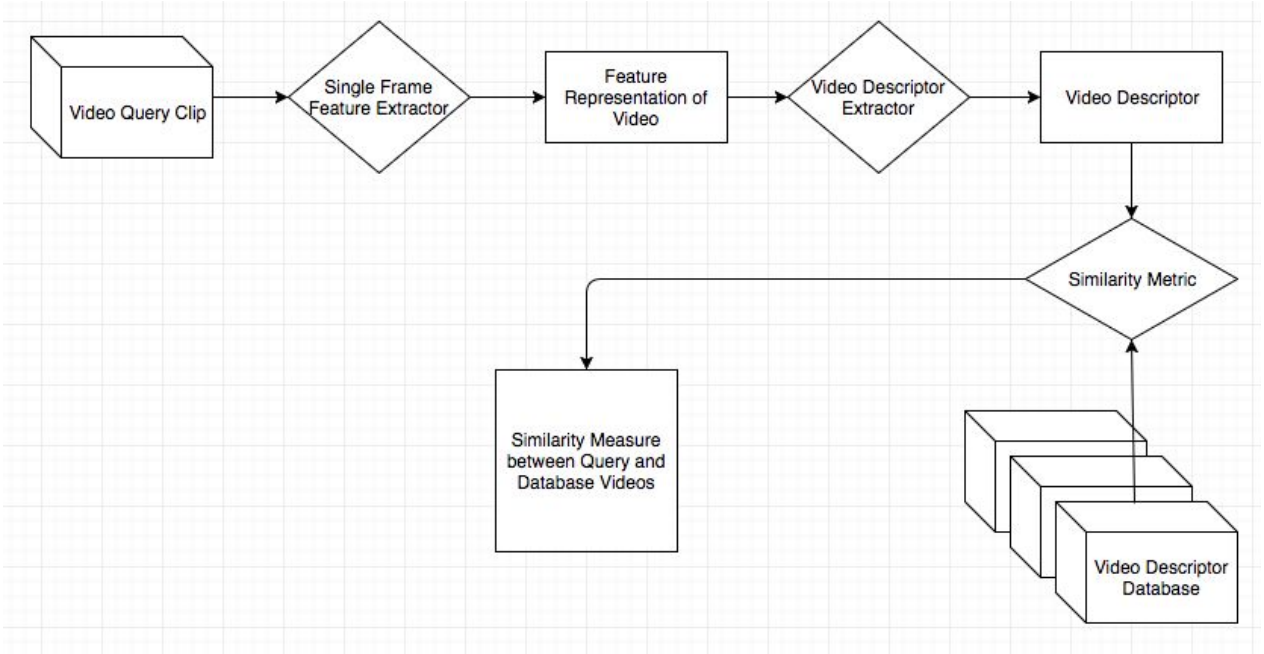
1. Analysis of results shows that the Spatio-temporal CNN approach performed well against prior work.
2. The architecture has few parameters and are quick to train.
3. Mentions portions where the method is weaker when compared to others.
4. Explanation of methods is organized and concise.

Cons:

1. Does not describe EndoTube dataset in detail, or show examples.
2. Current method does not perform well on more unregularized videos.
3. Best results require Tool Information, not always available.

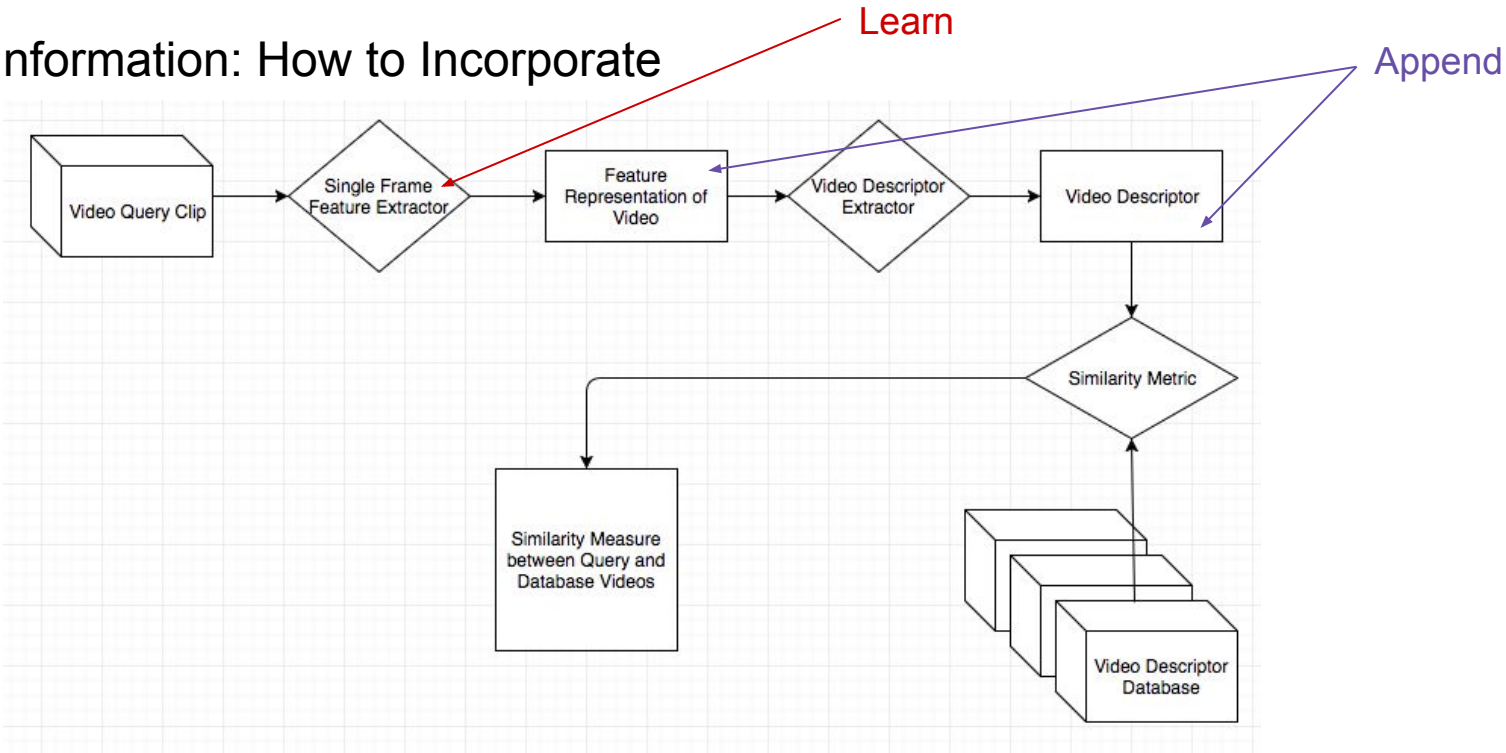
Most Important Relevance to Project

Tool Information: How to Incorporate



Most Important Relevance to Project

Tool Information: How to Incorporate



References

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

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

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