Real-time Surgical Tools Recognition in Total Knee Arthroplasty Using Deep Neural Networks

Moazzem Hossain[1,2], *Shoichi Nishio*[2], *Takafumi Hiranaka*[3] *and Syoji Kobashi*[2]

International University of Business Agriculture & Technology,[1]

University of Hyogo[2]

&

Takatsuki Hospital[3]

June 25, 2018

Outline

1. Introduction
   - Surgical Tools Recognition System.
   - Existing works.
   - Objectives.

2. Proposed Method.
   - Network Architecture.
   - Loss-function.
   - Bounding box prediction.
   - Dataset

3. Experimental Results

4. Conclusions

Surgical Tools Recognition

In order to accurately model a surgical process, a vital information needing to be retrieved is the presence of instruments in the surgical field of view.

Necessity of Surgery tools recognition systems

- Surgical workflow prediction,
- Surgical instruments monitoring,
- Surgery steps prediction,
- Robot-assisted surgery.

Literature review

Recent work:

- Localization of surgical tools in robot-assisted surgical training videos (Sarikaya et al., 2017)

  - Multimodal convolutional neural networks [1].

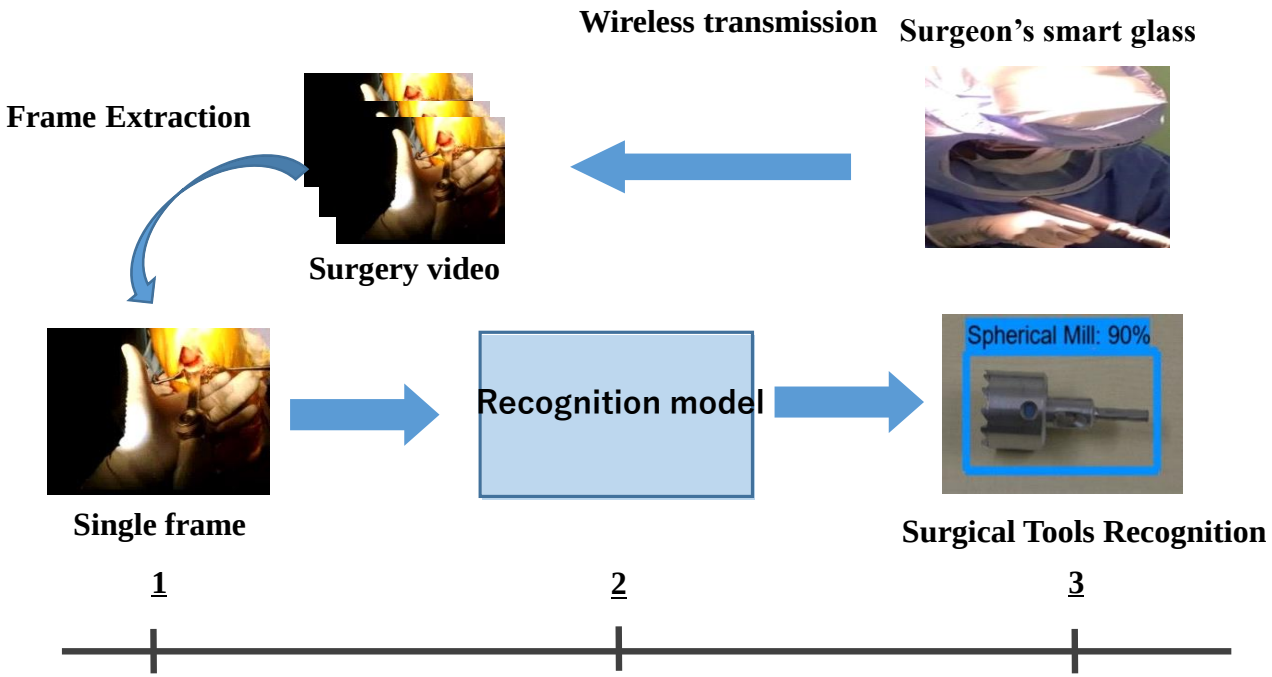
- Analyzes the specific suture and knots videos by activity, using the symbolic, texture and frequency characteristics (Zia et al., 2016).

  - Fast R-CNN [2].

1. D. Sarikaya, J. Corso, and G. Khurshid. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE Transactions on Medical Imaging, 36(7):1542–1549, 2017.
2. A. Zia, Y. Sharma, V. Bettadapura, E. Sarin, et al. Automated video based assessment of surgical skills for training and evaluation in medical schools. International Journal of Computer Assisted Radiology and Surgery, 11(9):1623–1636, 2016

Objectives

- To assist surgeon.
- To provide robust and accurate recognition.
- To identify surgical tools in Real-Time.



Proposed Method

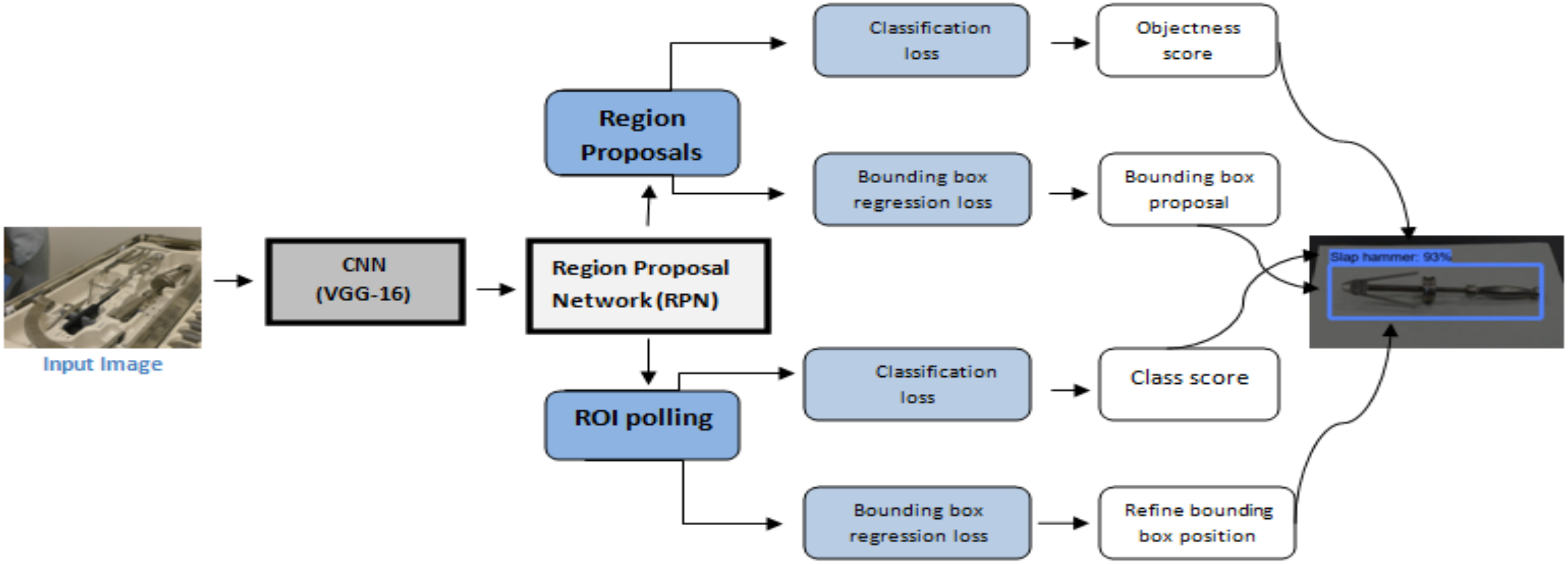


Figure 1: Our proposed method based on faster R-CNN architecture. The input to the network is a frame from a surgical video. The base network of Faster R-CNN is a VGG-16 convolutional neural network. This is connected to a region proposal network (RPN) that shares convolutional features with object detection networks. For each input image, the RPN generates region proposals that contain an object, and features are pooled over these regions before being passed to a final classification. The output is the bounding box positions of detected surgical tools.

Multi-Task Learning

- Rationales
  - Better bounding box prediction usually leads to better Surgery tools prediction.
  - Higher accuracy would also result in better bounding box prediction.

- Benefits
  - Accuracy of Multi-Task Learning is often better than Single-Task Learning.
  - Predicting different tasks simultaneously can save computational time.

Network Architecture

Table 2: Summary of parameters used by the proposed STD network. The convolutional block consists of convolution, batch normalization and activation layers.

Type of Layer	Filters	Size/Stride	Output Dimensions
Conv1 Max	16	3 × 3/1 2 × 2/2	416 × 416 × 16 208 × 208 × 16
Conv2 Max	32	3 × 3/1 2 × 2/2	208 × 208 × 32 104 × 104 × 32
Conv3 Max	64	3 × 3/1 2 × 2/2	104 × 104 × 64 52 × 52 × 64
Conv4 Max	128	3 × 3/1 2 × 2/2	52 × 52 × 128 26 × 26 × 128
Conv5 Max	256	3 × 3/1 2 × 2/2	26 × 26 × 256 13 × 13 × 256
Conv6 Max	512	3 × 3/1 2 × 2/1	13 × 13 × 512 13 × 13 × 512
Conv7 Conv8 Conv9	1024 1024 35	3 × 3/1 3 × 3/1 1 × 1/1	13 × 13 × 1024 13 × 13 × 1024 13 × 13 × 35
Detection			

Network Loss Function

i = anchor index in minibatch

Log loss

Ground truth objectness label

Smooth L1 loss

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

True box coordinates

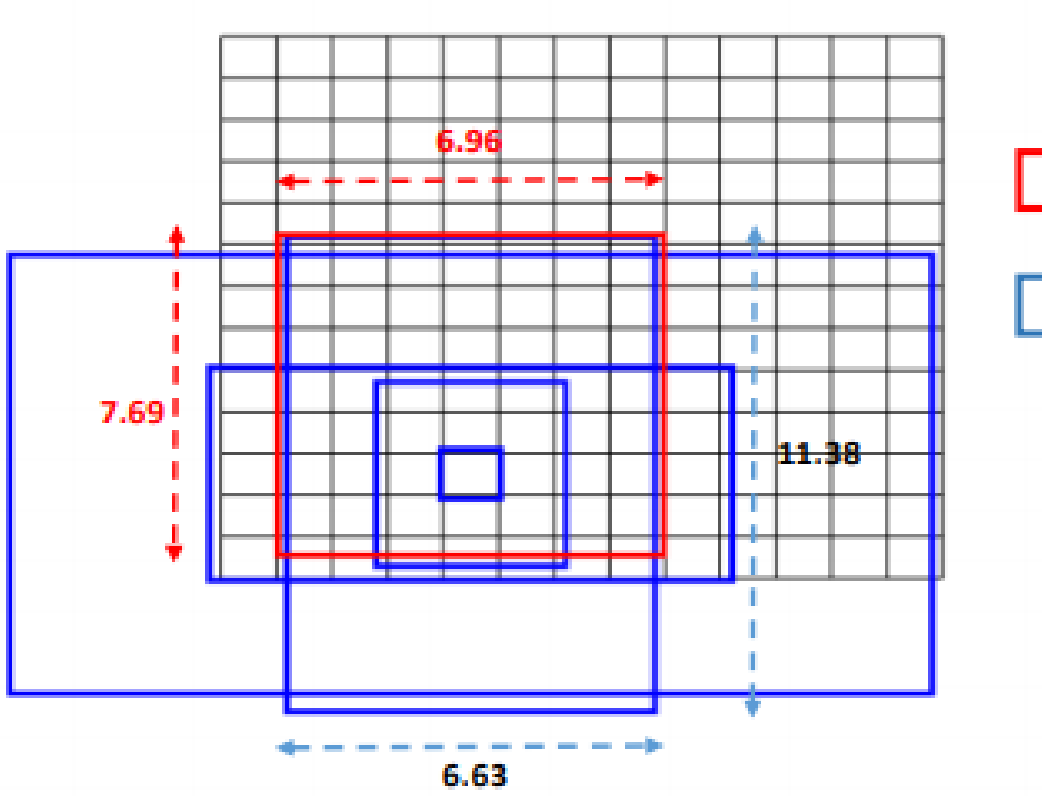
Coordinates of the predicted bounding box for anchor *i*

Predicted probability of being an object for anchor *i*

In practice $\lambda = 10$, so that both terms are roughly equally balanced

N_{cls} = Number of anchors in minibatch (~ 256)
N_{reg} = Number of anchor locations (~ 2400)

Bounding Box Prediction



Ground Truth: $box_{gt} = (x, y, w, h)$

Prediction: $box_{pd} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$

$$IoU = \frac{\cap\,(box_{gt}, box_{pd})}{\cup\,(box_{gt}, box_{pd})}$$

Figure 2: An illustration to convey the grid process of final output feature map. The blue boxes are the predicted bounding boxes that are computed based on the 5 pre-defined anchors.

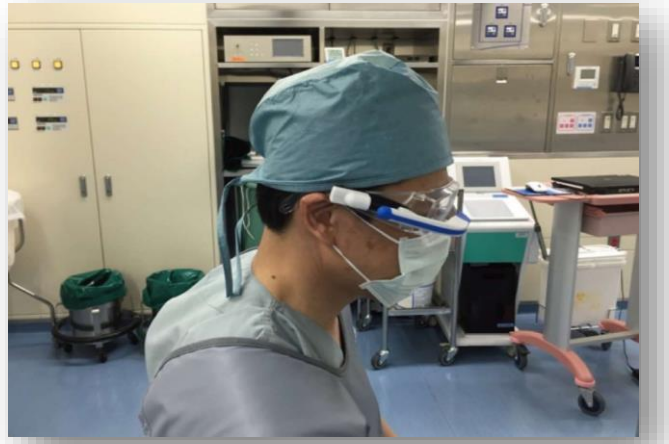
Training Protocol

- Training Procedure
  1. Manually annotating surgical tools bounding box regions from the training set.
  2. Normalize the annotations with respect to the width and height of image.
  3. Fine-tune on existing classification network (transfer learning).
- Training Parameters
  1. Batch size: 64
  2. Momentum= 0.9
  3. Decay=0.0005
  4. Learning rate=0.001 (steps policy)
  5. Max batches = 40200

Dataset

Our dataset generated by wearable smart glass consists of:

- Total number of videos: 16
- Total number of frames: 35,000
- 35,000 frames labelled with binary annotations indicating presence and absence of thirty one surgical tools.
- Leave one out cross validation



(a)

(b)

Experimental Dataset

(a)

(b)

(c)

(d)

(e)

(f)

Experimental Results

Methods	MAP %	Detection time	Min %	Max
Proposed method (Faster R-CNN)	88 (87.60)	0.075 s	75	96
Fast R-CNN	84 (84.48)	0.159 s	62	94
DPM	76 (76.00)	2.3 s	57	84
Edge+ Fast R-CNN	20	0.134	12	35

Experimental Results (2)

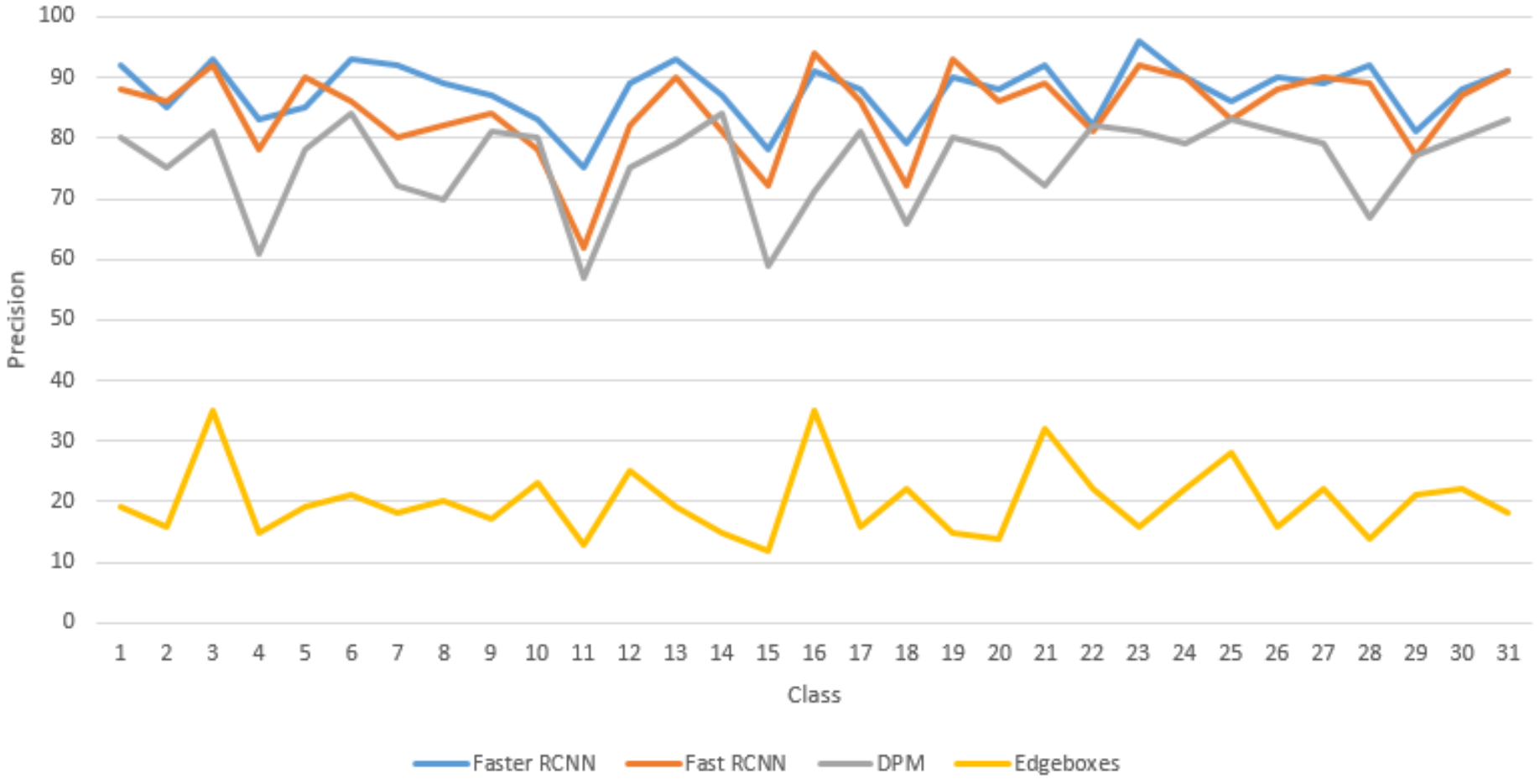


Figure 4: Our proposed method (Faster R-CNN) outperforms other methods.

Experimental Results (3)

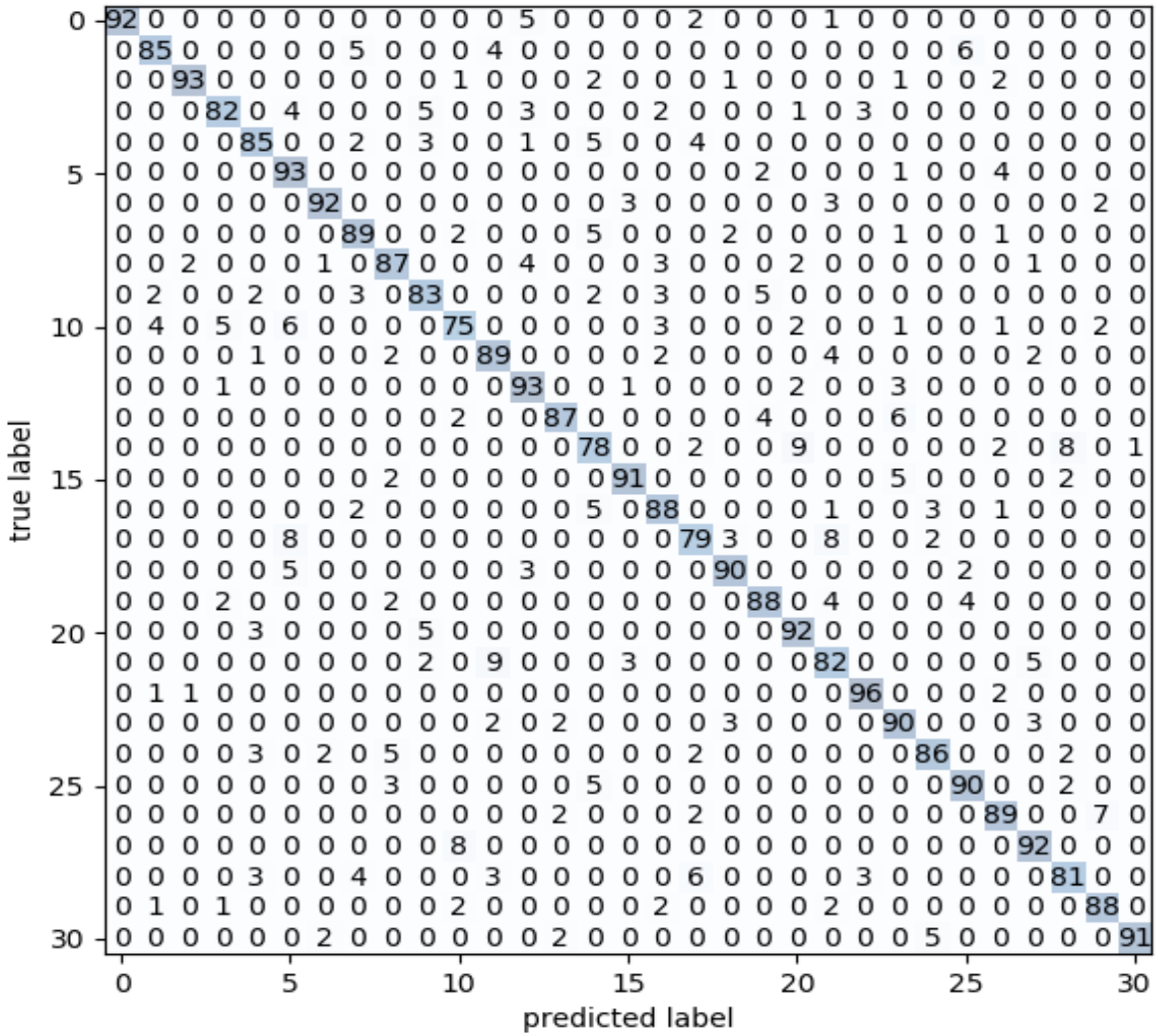
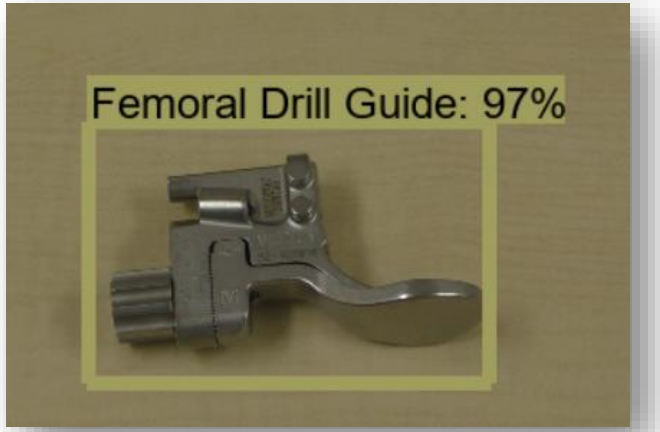
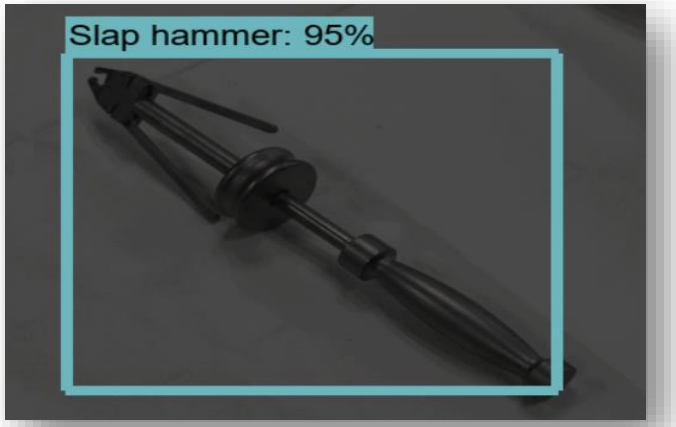


Figure 5: The confusion matrix of our proposed method indicate each tools true positive detection as well as some misclassification due to the similarity of the surgical tools.

Qualitative Results





Conclusions

- Our method outperforms previous work on frame-level presence detection in real-world surgical videos, and runs in real-time.
- Our work can be used to extract rich surgical assessment metrics such as tool usage patterns, movement range, and economy of motion.
- Further refinement of the proposed method will focus on dealing with the conditions of occlusion and multi-instruments, which is planned to be achieved by constructing a more efficient and robust line feature detector and a more flexible CNN structure.

Thank You