Development and Validation of a Model for Laparoscopic Colorectal Surgical Instrument Recognition Using Convolutional Neural Network-Based Instance Segmentation and Videos of Laparoscopic Procedures

Daichi Kitaguchi, MD; Younae Lee, MS; Kazuyuki Hayashi, MS; Kei Nakajima, MD; Shigehiro Kojima, MD, PhD; Hiro Hasegawa, MD, PhD; Nobuyoshi Takeshita, MD, PhD; Kensaku Mori, PhD; Masaaki Ito, MD, PhD

Abstract

IMPORTANCE Deep learning-based automatic surgical instrument recognition is an indispensable technology for surgical research and development. However, pixel-level recognition with high accuracy is required to make it suitable for surgical automation.

OBJECTIVE To develop a deep learning model that can simultaneously recognize 8 types of surgical instruments frequently used in laparoscopic colorectal operations and evaluate its recognition performance.

DESIGN, SETTING, AND PARTICIPANTS This quality improvement study was conducted at a single institution with a multi-institutional data set. Laparoscopic colorectal surgical videos recorded between April 1, 2009, and December 31, 2021, were included in the video data set. Deep learning-based instance segmentation, an image recognition approach that recognizes each object individually and pixel by pixel instead of roughly enclosing with a bounding box, was performed for 8 types of surgical instruments.

MAIN OUTCOMES AND MEASURES Average precision, calculated from the area under the precision-recall curve, was used as an evaluation metric. The average precision represents the number of instances of true-positive, false-positive, and false-negative results, and the mean average precision value for 8 types of surgical instruments was calculated. Five-fold cross-validation was used as the validation method. The annotation data set was split into 5 segments, of which 4 were used for training and the remainder for validation. The data set was split at the per-case level instead of the per-frame level; thus, the images extracted from an intraoperative video in the training set never appeared in the validation set. Validation was performed for all 5 validation sets, and the average mean average precision was calculated.

RESULTS In total, 337 laparoscopic colorectal surgical videos were used. Pixel-by-pixel annotation was manually performed for 81,760 labels on 38,628 static images, constituting the annotation data set. The mean average precisions of the instance segmentation for surgical instruments were 90.9% for 3 instruments, 90.3% for 4 instruments, 91.6% for 6 instruments, and 91.8% for 8 instruments.

CONCLUSIONS AND RELEVANCE A deep learning-based instance segmentation model that simultaneously recognizes 8 types of surgical instruments with high accuracy was successfully developed. The accuracy was maintained even when the number of types of surgical instruments increased. This model can be applied to surgical innovations, such as intraoperative navigation and surgical automation.

Key Points

Question Is an automatic surgical instrument recognition model with wide applicability to multiple types of instruments and pixel-level high recognition accuracy feasible?

Findings This quality improvement study for development and validation used a multi-institutional data set that consisted of 337 laparoscopic colorectal surgical videos. The mean average precisions of the instance segmentation of surgical instruments were 90.9% for 3 instruments, 90.3% for 4 instruments, 91.6% for 6 instruments, and 91.8% for 8 instruments.

Meaning This study suggests that deep learning can be used to simultaneously recognize multiple types of instruments with high accuracy.
Introduction

Minimally invasive surgery (MIS), such as laparoscopic and robotic types, continues to play an important role in general surgery as an alternative to traditional open surgery. Since the 1980s, technological advancement and innovation have led to a rapid increase in the number of MIS surgical techniques, because they are viewed as more desirable. Minimally invasive surgery has better cosmetic results than traditional open surgery and has become increasingly common during the past years. On the other hand, the reliance on a surgeon’s sense of touch when performing surgical procedures has waned, and surgeons have become highly dependent on the visual information provided by laparoscopic monitors. To be more dependable, laparoscopic cameras must be able to rapidly capture and track the surgical instrument being used so that unsafe blind surgical procedures can be avoided. Therefore, a prerequisite in MIS is the ability to accurately recognize the location of surgical instruments and track them; in other words, we can extract much of the information associated with the surgery from such kinematic data.

In recent years, computer-based automation, which self-captures and self-analyzes large amounts and a wide variety of data—exemplified by self-driving cars—has been advancing and rapidly pervading all aspects of daily life. Artificial intelligence, especially its subdomain of deep learning, is used to make big data interpretable and usable. One area of artificial intelligence that has particularly bourgeoned during the last decade is computer vision (CV), which is an interdisciplinary scientific field that deals with how computers gain a high-level understanding of digital images or videos and use the knowledge to perform functions, such as object identification, tracking, and scene recognition.

Deep learning-based CV technology has been applied to various tasks in laparoscopic surgery, such as anatomical structures, surgical action, and step or phase recognition. Applying this technology to surgical instrument recognition can enable the development of robotic camera holders with automated real-time instrument tracking during surgery, and it can be used for surgical skill assessment, augmented reality, and depth enhancement during MIS. Although the goal of computer-assisted surgery is full automation, it must be realized in stages (e.g., starting with intestinal anastomosis). In any stage, automatic surgical instrument recognition using deep learning-based CV is an indispensable fundamental technology. Therefore, an automatic surgical instrument recognition model with wide applicability to multiple types of instruments and pixel-level high recognition accuracy is required for research and development in the field of computer-assisted surgery.

The objectives of this study were to develop a deep learning model that can simultaneously recognize 8 types of surgical instruments frequently used in laparoscopic colorectal operations and to evaluate its recognition performance. In addition, the applicability of the recognition results of the developed model to surgical progress monitoring was examined.

Methods

Study Design

This quality improvement study for development and validation was conducted at a single institution using a multi-institutional intraoperative video data set. This study followed the reporting guidelines of the Standards for Quality Improvement Reporting Excellence (SQUIRE) and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines. The study protocol was reviewed and approved by the Ethics Committee of the National Cancer Center Hospital East, Chiba, Japan. Informed consent was obtained from patients in the form of an opt-out on the study website, and data from those who rejected participation were excluded. The study conformed to the provisions of the Declaration of Helsinki in 1964 (and revised in Brazil in 2013).
Annotation Data Set

Laparoscopic colorectal surgical videos recorded between April 1, 2009, and December 31, 2021, were included in the video data set in this study. Every intraoperative video was converted to the MP4 video format with a display resolution of 1280 × 720 pixels and frame rate of 30 frames per second. A total of 38,628 static images capturing the surgical instruments to be recognized were randomly extracted from the video data set and included in the annotation data set for this study.

The following 8 types of surgical instruments were used for recognition: surgical shears (Harmonic HD 1000i and Harmonic ACE, Ethicon Inc) (SI-1), spatula-type electrode (Olympus Co Ltd) (SI-2), atraumatic universal forceps with grooves (Aesculap AdTec, B Braun AG) (SI-3), dissection forceps (HiQ+ Maryland Dissection Forceps, Olympus Co Ltd) (SI-4), endoscopic clip appliers (Ligamax 5 Endoscopic Clip Applier, Ethicon Inc, and Endo Clip III 5-mm Clip Applier, Medtronic Plc) (SI-5), staplers (Echelon Stapler, Ethicon Inc, and Endo GIA Reload with Tri-Staple Technology, Medtronic Plc) (SI-6), grasping forceps (Croce-Olmi grasping forceps, Karl Storz SE & Co KG) (SI-7), and a suction/irrigation system (HiQ+ suction/irrigation system, Olympus Co Ltd) (SI-8). Reference images of these 8 types of surgical instruments are shown in Figure 1A.

Annotation was performed by 29 individuals who were not physicians under the supervision of 2 board-certified surgeons (D.K. and H.H.), and 8 types of annotation labels were manually assigned pixel by pixel by drawing directly on the area of each surgical instrument on the static images using digital drawing pens (Wacom Cintiq Pro and Wacom Pro Pen 2, Wacom Co Ltd, or Microsoft Surface Pro 7 and Microsoft Surface Pen, Microsoft Corp) equipment. Reference images of the annotations are shown in Figure 1B.

Figure 1. Reference Images and Annotations of the 8 Types of Surgical Instruments

Panel A shows reference images of the 8 types of surgical instruments. Panel B shows reference images of the surgical instrument annotation.
CNN-Based Instance Segmentation

Convolutional neural network (CNN)-based instance segmentation was performed with Mask R-CNN as the architecture network and ResNet-50 as the backbone network. Instance segmentation is an image recognition approach that recognizes each object individually and pixel by pixel, instead of enclosing all objects within a bounding box. The network weight was initialized to a pretrained one on the Microsoft Common Objects in Context data set, which is a large-scale data set for object detection, segmentation, and captioning that contains 2,500,000 labeled instances in 328,000 images for 91 common object categories. Every annotated image in the training set was input into the deep learning model, and fine-tuning was then performed for the training set. The best epoch model, based on the model performance on the validation set, was selected. For data augmentation, the resize, random horizontal flip, random vertical flip, and normalize functions were used.

Code and Computer Specifications

The code was written in Python, version 3.6 (Python Software Foundation) and is available via GitHub on reasonable request. The implementation of the model was based on MMDetection, which is an open-source Python library for object detection and instance segmentation. A computer equipped with a Tesla T4 GPU with 16 GB of VRAM (NVIDIA) and a Xeon CPU Gold 6230 (2.10 GHz, 20 Core) with 48 GB of RAM (Intel Corp) was used for network training.

Validation and Evaluation

Average precision (AP) was used as an evaluation metric for the CNN-based surgical instrument recognition task in this study. Average precision is calculated from the area under the precision-recall curve, which is described based on the number of true-positive (TP), false-positive (FP), and false-negative (FN) results. Intersection over union (IoU) is a metric that signifies the degree of overlap between the area annotated as the ground truth (A) and the predicted area output by the CNN (B); an IoU of 1 signifies that there is perfect overlap for the corresponding pair, A and B, whereas an IoU of 0 signifies that there is no overlap. An IoU reference image is shown in Figure 2A.

In this study, the threshold value of IoU was set to 0.75; that is, when the IoU of the corresponding pair of A and B was more than 0.75, it was defined as TP; when the IoU was less than 0.75, it was defined as FN; and when there was no A corresponding to B, it was defined as FP. The formulas for precision, recall, and IoU are as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \quad \text{IoU} = \frac{A \cap B}{A \cup B^c}
\]

The mean AP (mAP) metric is widely used in object detection and instance segmentation tasks. In this study, mAP denotes the mean AP value for 8 types of surgical instruments.

Figure 2. Illustration of Intersection Over Union and 5-Fold Cross-validation

In 5-fold cross-validation, the annotation data set is split into 5 segments, of which 4 are used as training sets and the remainder as the validation set. The validation is performed for all 5 validation sets, and the mean value is calculated as the performance metric.
Five-fold cross-validation and hold-out validation were used as the validation methods in this study as appropriate. In 5-fold cross-validation, the annotation data set was split into 5 segments, of which 4 were used as training sets and the remainder as the validation set. The data split was performed at the per-case level instead of the per-frame level; thus, the images extracted from an intraoperative video in the training set never appeared in the validation set. The validation was performed for all 5 validation sets, and the average mAP was calculated. A reference image of the 5-fold cross-validation is shown in Figure 2B. In addition, a misrecognition pattern analysis was performed.

**Automatic Surgical Progress Monitoring**

In colorectal surgery, transection of the inferior mesenteric artery (IMA) and rectum is an important landmark for monitoring surgical progress. In the standard IMA root transection procedure, the dorsal side of the IMA root is dissected with SI-4 and clipped with SI-5; SI-6 is only used for rectal transections.

To examine the applicability of the proposed model to an automatic surgical progress monitoring system, we used the results of the surgical instrument recognition on a new intraoperative video of laparoscopic colorectal surgery. Specifically, we evaluated whether continuous recognition of SI-4 followed by continuous recognition of SI-5 and the timing of continuous recognition of SI-6 were included in the previously defined surgical step of IMA and rectum transections, respectively.

**Results**

**Annotation Data Set**

In total, 337 laparoscopic colorectal surgical videos were used in this study, and 81,760 labels were annotated for 38,628 static images extracted from the video data set. The distribution of the number of annotation labels for each surgical instrument is shown in Figure 3A: 5694 for SI-1, 4358 for SI-2, 41,778 for SI-3, 5248 for SI-4, 5254 for SI-5, 5810 for SI-6, 7022 for SI-7, and 6596 for SI-8. SI-3 had numerous annotation labels because up to 4 could appear in a single static image, and they frequently appeared with other surgical instruments.

**Instance Segmentation Accuracy**

The average (SD) mAP of instance segmentation for the 8 types of surgical instruments obtained from 5-fold cross-validation was 91.9% (0.34%). In this study, the threshold value of IoU was set to 0.75, but as supplemental data, the mAP values at multiple thresholds ranging from 0.5 to 0.95 are plotted and shown in the eFigure in the Supplement.
The change in the mAP when the number of types of surgical instruments to be recognized was increased is shown in Figure 3B. The mAPs of instance segmentation were 90.9% for 3 surgical instruments, 91.6% for 4 surgical instruments, and 91.8% for 8 surgical instruments.

The AP of instance segmentation for each surgical instrument is shown in Figure 3C: 91.4% for SI-1, 95.6% for SI-2, 85.4% for SI-3, 90.7% for SI-4, 96.4% for SI-5, 91.2% for SI-6, 85.6% for SI-7, and 95.7% for SI-8. The recognition results of instance segmentation for the 8 types of surgical instruments in a complete laparoscopic sigmoid colon resection procedure are shown in Figure 4 and the Video.

### Misrecognition Pattern Analysis

The misrecognition pattern obtained when instance segmentation was performed on 1 validation set is shown in the confusion matrix of Figure 5A; 213 misrecognitions were observed in the validation set. Intracorporeal artifacts other than surgical instruments (eg, anvil shaft of a circular stapler, drainage tube, and trocar) tended to be misrecognized as SI-3, which was the main cause of misrecognition (125 of 213 misrecognitions [58.7%]). Representative images of the misrecognition patterns are shown in Figure 5B.

The vertical axis indicates each surgical instrument, the horizontal axis is the operating time, and the timing of the appearance of each surgical instrument is plotted with black lines to show the time series. The first and second outlined areas represent the surgical steps of inferior mesenteric artery (IMA) transection and rectal transection, respectively. The third and fourth outlined areas indicate extracorporeal procedures and represent the surgical steps of specimen extraction and wound closure, respectively.
Automatic Surgical Progress Monitoring

As shown in Figure 4, the continuous recognition of SI-4 followed by the continuous recognition of SI-5 was included in the surgical step of the IMA transection. In addition, the timing of the continuous recognition of SI-6 was included in the surgical step of rectal transection.

Discussion

In this quality improvement study, we successfully developed a CNN-based instance segmentation model that can simultaneously recognize 8 types of surgical instruments with high recognition accuracy. Almost all surgical instruments used in laparoscopic colorectal operations are covered by these 8 types; therefore, the developed model can be applied to intraoperative navigation for a series of laparoscopic colorectal surgical procedures.

There are several approaches to CNN-based CV for image recognition tasks. The bounding box-based object detection approach has often been used\(^{22,23}\); however, in future developments in the surgical research field, such as the automation of surgery, more detailed pixel-by-pixel recognition information will be essential, instead of bounding box-based object detection.

Semantic segmentation is a CNN-based CV approach that enables pixel-by-pixel image recognition. Recently, it has also been actively used in the research field of CV in surgery.\(^{24,25}\) Semantic segmentation attempts to specifically understand the role of each pixel in an image through the process of dividing whole images into pixel groupings that can then be labeled and classified. The boundaries of each object can be delineated; therefore, dense pixel-based predictions can be achieved. However, semantic segmentation cannot distinguish them individually when objects of the same type overlap. Instance segmentation overcomes this limitation because it can distinguish overlapping objects, even when they are of the same type. It enables the visualization of complicated images with multiple overlapping objects and identifies their relationships. In this study, we selected instance segmentation as the image recognition approach because the crossing of multiple surgical instruments during surgery is a common occurrence.

In general, multiclass recognition is difficult—the more types of objects there are to be recognized, the more difficult the task becomes. However, the accuracy of the recognition task for the 3, 4, 6, and 8 types of surgical instruments was almost the same in this study (Figure 3B). One possible reason for this is that we prepared a sufficient amount of annotated data. These data comprised more than 80,000 annotation labels for the training of the CNN-based instant segmentation model, enabling us to overcome the difficulties besetting the multiclass recognition task.

Although the recognition accuracy was quite high, several misrecognition patterns needed to be measured, which were mainly caused by intracorporeal artifacts other than surgical instruments in this study (Figure 5). However, we believe that this problem can be solved in the future by intentionally adding annotation images that capture such artifacts and training the model such that they are part of the background. Otherwise, such artifacts can be added to the recognition target with annotations. In addition, although the situation that the surgical instrument tip was out of view was also expected to cause misrecognition, this pattern was minor in this study. This pattern was thought to be because such situations were rare during surgery to begin with, and much of the information on shafts and tips was used in the predictions. Differential identification among surgical instruments with the same shaft and different tip options is a challenge that will be tackled in future work.

The proposed instance segmentation model for surgical instrument recognition has various applications in surgical research and developing fields, such as automated robotic camera holders and automatic surgical skill assessments based on surgical instrument tracking. This study demonstrated the feasibility of applying the model to automatic surgical progress monitoring.

Accurately monitoring surgical progress in real time for optimal operating room (OR) logistics is necessary.\(^{26}\) This monitoring enables OR resource optimization, and the spare time and human
resources lead to patient comfort and safety. In addition, if a certain surgical step takes longer than planned and causes delays, this may indicate that some kind of intraoperative adverse event has occurred. Automatic self-extracting, analyzing, and monitoring systems for surgical progress enable the entire OR staff to easily share crisis awareness, which leads to improved OR team performance in terms of quick decisions and actions.

In this study, we focused on the appearance information of SI-4/5 and SI-6 in the surgical steps of IMA transection and rectal transection in laparoscopic colorectal surgery, respectively; however, other surgical instruments appear in specific surgical steps in every type of surgery. Therefore, the proposed approach for automatic surgical progress monitoring based on surgical instrument recognition is promising because of its high applicability. In the future, it may be possible to monitor not only the progress of the surgery but also the event of intraoperative bleeding based on the appearance information of surgical instruments used for hemostasis, such as SI-8.

Limitations
This study has several limitations. First, although the proposed instance segmentation model for surgical instrument recognition has various applications in surgical research and development, only a proof-of-concept evaluation was performed; the model was not applied to actual surgical practice. Second, only validation was conducted; tests using data from an external cohort were not performed. However, the annotation data set was sufficiently large, and stable recognition accuracy was obtained from 5-fold cross-validation; therefore, the generalization performance was high. Third, for automatic surgical progress monitoring, only preliminary results were obtained from a single intraoperative video. For future validation, we must determine the minimum frame rate for continuous recognition. Fourth, because SI-3 tended to appear with complex intersections with other surgical instruments, the corresponding mAP was lower than for the other instruments, despite having more annotation labels. Improving the recognition accuracy of SI-3 is also a future task.

Conclusions
We developed an instance segmentation model that can simultaneously recognize 8 types of surgical instruments that are frequently used in laparoscopic colorectal operations. In this quality improvement study, the recognition accuracy of the model was high and was maintained even when the number of types of surgical instruments to be recognized increased. Surgical instrument recognition is an important fundamental technology for various surgical research and development areas. The proposed model can be applied not only to the proposed automatic surgical progress monitoring in this study but also for future computer-assisted surgery realization or surgical automation.
Author Contributions: Dr Ito had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Kitaguchi, Hayashi, Nakajima, Kojima, Hasegawa, Takeshita, Mori, Ito.

Acquisition, analysis, or interpretation of data: Kitaguchi, Lee, Hayashi, Nakajima, Mori, Ito.

Drafting of the manuscript: Kitaguchi, Lee, Nakajima, Mori.

Critical revision of the manuscript for important intellectual content: Hayashi, Kojima, Hasegawa, Takeshita, Ito.

Statistical analysis: Kitaguchi, Lee, Nakajima.

Obtained funding: Takeshita, Ito.

Administrative, technical, or material support: Hayashi, Kojima, Hasegawa, Takeshita, Mori, Ito.

Supervision: Hayashi, Takeshita, Mori, Ito.

Conflict of Interest Disclosures: Dr Takeshita reported receiving grants from the Japan Agency for Medical Research and Development during the conduct of the study. Dr Mori reported receiving grants from the Japan Agency for Medical Research and Development, Olympus, and Japan Society for the Promotion of Science during the conduct of the study and grants and royalties from Cybernet and Morita Mfg outside the submitted work. No other disclosures were reported.

Funding/Support: This research was funded by grant JP20he2302003h0102 (Dr Ito) from the Japan Agency for Medical Research and Development.

Role of the Funder/Sponsor: The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES
Evaluation of a Model for Laparoscopic Colorectal Surgical Instrument Recognition


SUPPLEMENT.
eFigure. Mean Average Precision Values at Multiple Thresholds Ranging From 0.5 to 0.95