1版

様　式　C－１９、F－１９－１、Z－１９　（共通）

科研費
KAKENHI

科学研究費助成事業　　研究成果報告書

令和　４ 年　６ 月　１ 日現在

機関番号：１４４０１

研究種目：研究活動スタート支援

研究期間：2020～2021

課題番号：２０Ｋ２３３４３

研究課題名（和文）Removing the Burden of Data Labeling: Automatic Surgical Video Understanding with Unsupervised Learning

研究課題名（英文）Removing the Burden of Data Labeling: Automatic Surgical Video Understanding with Unsupervised Learning

研究代表者

李　良知（LI, LIANGZHI）

大阪大学・データビリティフロンティア機構・特任助教（常勤）

研究者番号：１０８７５５４５

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：既存の手術映像自動解析モデルの多くは、人によるラベル付けデータを必要とする。本プロジェクトでは、より少ない人手で空間および時間のセグメンテーションを行うための学習手法を設計することを目的としている。主に以下のようなサブテーマについて研究した。
1.動画像のセマンティックセグメンテーションのための新しい学習手法を開発し、非ラベル化データを利用してセグメンテーションのパフォーマンスを向上させた。
2.手術映像の時間分割と時間解析モデル。手術時間を自動的に予測するための検索に基づく手法を開発した。

研究成果の学術的意義や社会的意義
The establishment of techniques capable of understanding what, where, and when is happening with no requirements on human efforts will ease the task of surgery indexing, clinical training, etc.

研究成果の概要（英文）：As most of the existing automatic surgical video analysis models require a large number of manually labeled data for training, this project aims to design a learning method to perform spatial and temporal segmentations with smaller requirements of humans’ input. During this project, I mainly studied the following sub-topics towards the goal.
1.Surgical images/frames semantic segmentation in a weakly-supervised way. I developed a new training strategy for video semantic segmentation models to utilized unlabeled data to improve their segmentation performance.
2.Surgical videos temporal analysis using no labels. I developed a retrieval-based method to automatically predict surgical duration.

研究分野：Computer Vision

キーワード：Few-shot Learning　Semantic Segmentation　Video Understanding　Medical Images　Computer Vision　Surgical Analysis　Deep Learning

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

Well-analyzed surgical videos are of great value for not only training clinical students but also helping professionals. For that purpose, there have been lots of works on the **automatic analysis of surgical videos**. Encouraging results have been achieved in the understanding of the spatial information (what and where are the objects) as well as the temporal information (what and when is happening). However, it remains **a challenging unsolved problem** when no enough ground-truth labels are available, as most of the existing models belong to *supervised learning*, which needs labels to teach the deep learning models, as shown in Figure 1. This fact has extremely limited the broad application of the automatic surgical video analysis system. Designing a new computer vision (CV) method which can remove the labeling burden is **the first step to enable low-cost automatic surgery analysis**. Therefore, the key scientific question this project aims to answer is, how to extract spatial and temporal information from surgical videos with no labeled data available.

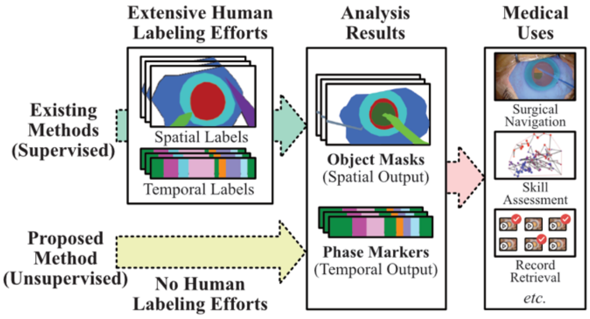

Extensive Human Labeling Efforts
Analysis Results
Medical Uses
Existing Methods (Supervised)
Spatial Labels
Temporal Labels
Object Masks (Spatial Output)
Phase Markers (Temporal Output)
Surgical Navigation
Skill Assessment
Record Retrieval
etc.
Proposed Method (Unsupervised)
No Human Labeling Efforts

Figure 1. Supervised and unsupervised methods

２．研究の目的

The objective of this research is to make some progress towards enabling the unsupervised spatial understanding (USU) and unsupervised temporal understanding (UTU) for surgical videos. As shown in Figure 2, we will first design a new USU method using hidden temporal information (*local consistency* for generating object proposals, *i.e.*, finding the pixels which always move together and thus may be an independent object, and *motion clues* for differentiating and merging proposals into object masks, *i.e.*, which proposals should be same object). Then we will address the existing problems in UTU methods with the obtained spatial information (*object knowledge* for introducing more temporal differences, *i.e.*, which instruments or tissues are involved). The research objectives (RO) include:

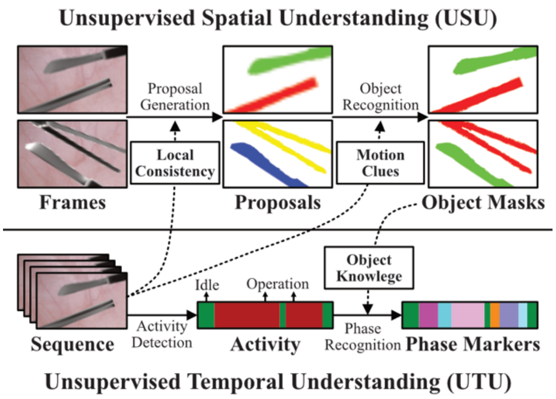


Figure 2. The proposed unsupervised method

- **RO1**: Enable the USU of surgery scenes using temporal clues.
- **RO2**: Utilize spatial knowledge to improve UTU performance.

３．研究の方法

There are some attempts in CV aiming at performing video analysis without ground-truth labels. These *unsupervised learning* methods **need no human-made labels**. However, few researchers have adopted them in the medical area, as the objects in medical images are either too similar to differentiate, which is a key obstacle to USU, or too small to cause significant changes which are required for UTU. In fact, medical images do have some inherent information to make *unsupervised learning* possible. For example, the objects in medical images usually have unique movements which are much easier to recognize, *e.g.*, some forceps and scissors may have similar appearances but with different movements (grasping/cutting); the exists of some objects may also serve as a sign to infer the possible surgical phases because using different combinations of instruments usually means different surgical phases. This kind of spatial-temporal information is pervasive among various surgery types and can help to enable automatic surgery understanding and **has not yet been studied**.

Therefore, to make progress towards enabling automatic surgery video analysis, this project is divided into three modules.

- **M1**: We will design a new USU model for object segmentation.
- **M2**: We will propose a new UTU model for phase segmentation.
- **M3**: We plan to develop an automatic segmentation software to output spatial and temporal results.

The first 12 months will be equally and sequentially dedicated to M1 and M2, respectively for the object and phase segmentation. We will make prototype implementations and test their performance. For *the last 6 months* in M3, we will develop software for surgery analysis using the proposed method. Also, we are going to submit two journal/conference papers to report our work in M1 and M2.

４．研究成果

(1) Weakly supervised semantic segmentation

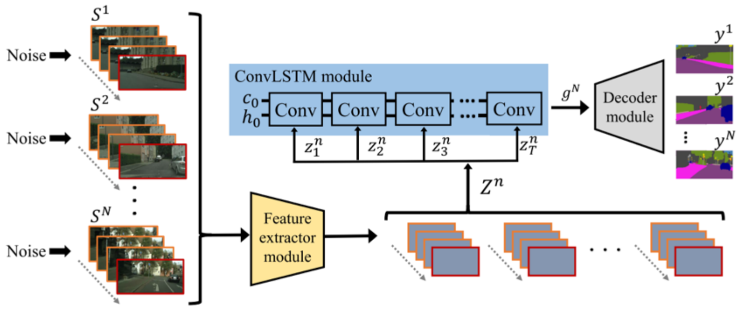


Figure 3. The Noisy-LSTM model.

Semantic video segmentation is a key challenge for various applications. As shown in Figure 3, I present a new model named Noisy-LSTM, which is trainable in an end-to-end manner, with convolutional LSTMs (ConvLSTMs) to leverage the temporal coherence in video frames, together with a simple yet effective training strategy that replaces a frame in a given video sequence with noises. The training strategy spoils the temporal coherence in video frames and thus makes the temporal links in ConvLSTMs unreliable; this may consequently improve the ability of the model to extract features from video frames and serve as a regularizer to avoid overfitting, without requiring extra data annotations or computational costs. Experimental results demonstrate that the proposed model can achieve state-of-the-art performances on both the CityScapes and EndoVis2018 (as shown in Figure 4) datasets.

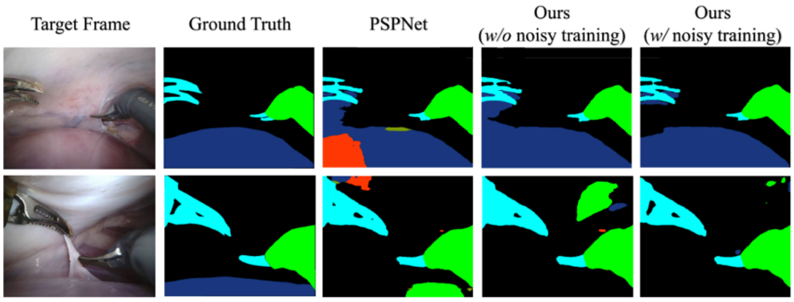


Figure 4. Segmentation performance on surgical videos.

(2) Video temporal segmentation and duration estimation

Estimating the surgery length has the potential to be utilized as skill assessment, surgical training, or efficient surgical facility utilization especially if it is done in real-time as a remaining surgery duration (RSD). Surgical length reflects a certain level of efficiency and mastery of the surgeon in a well-standardized surgery such as cataract surgery. Real-time RSD estimation can enable optimization of operating room scheduling as well. In this paper, we design and develop a real-time RSD estimation method for cataract surgery that does not require manual labeling.

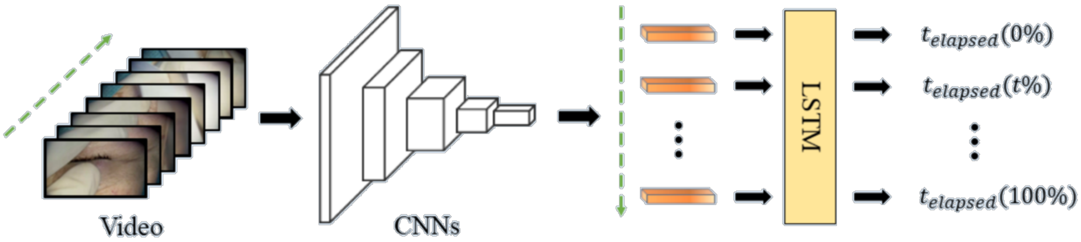


Figure 5. Surgery temporal segmentation and duration estimation model.

As shown in Figure 5, a regression method consisting of convolutional neural networks (CNNs) and long short-term memory (LSTM) is designed for RSD estimation. The model is firstly trained and evaluated for the single main target surgeon. Then, the fine-tuning strategy is used to transfer the model to the data of the other two surgeons. Mean average error (MAE in seconds) was used to evaluate the performance of the RSD estimation. The proposed method is compared with the naïve method which is based on the statistic of the historical data. A transferability experiment is also set to demonstrate the generalizability of the method.

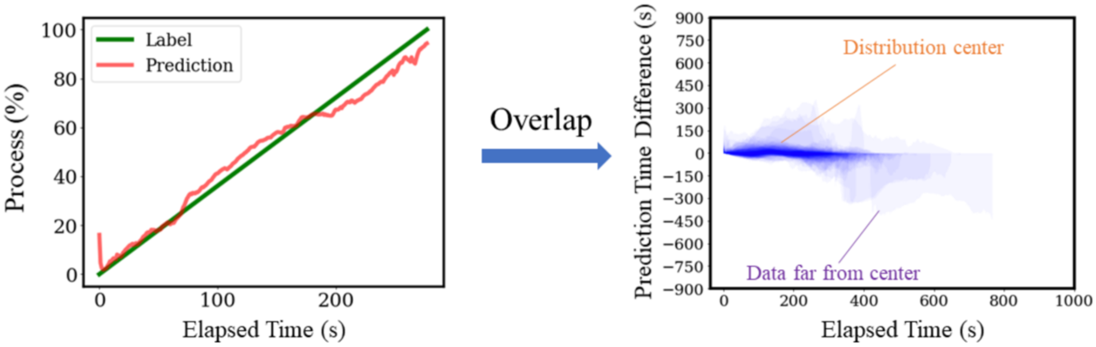


Figure 6. Surgery duration estimation performance.

As shown in Figure 6, the model shows low prediction error. The mean surgical time for the sample videos was 318.7 seconds (s) (standard deviation 83.4 s) for the main surgeon for the initial training. In our experiments, the lowest MAE of 19.4s (equal to about 6.4% of the mean surgical time) is achieved by our best-trained model for the independent test data of the main target surgeon. It reduces the MAE by 35.5s (-10.2%) compared to the naïve method. The fine-tuning strategy transfers the model trained for the main target to the data of other surgeons with only a small number of training data (20% of the pre-training). The MAEs for the other two surgeons are 28.3s and 30.6s with the fine-tuning model, which decreased by -8.1s and -7.5s than the Per-surgeon model (average declining of -7.8 s and 1.3% of video duration).

(3) Few-shot Learning

Few-shot learning (FSL) approaches, mostly neural network-based, are assuming that the pre-trained knowledge can be obtained from base (seen) categories and transferred to novel (unseen) categories. However, the black-box nature of neural networks makes it difficult to understand what is actually transferred, which may hamper its application in some risk-sensitive areas. As shown in Figure 7, I reveal a new way (MTUNet) to perform explainable FSL for image classification, using discriminative patterns and pairwise matching. MTUNet learns discriminative patterns for a given set of images of the base categories as shown in Figure 7 and uses all these patterns to represent both support and query images. With this representation, pairwise matching scores are computed among the support and query images, based on which the prediction for the query image is done. Both the patterns and the overall representation can be easily visualized to reveal the reason for the matching scores. Experimental results prove that the proposed method can achieve satisfactory explainability (as shown in Figure 8).

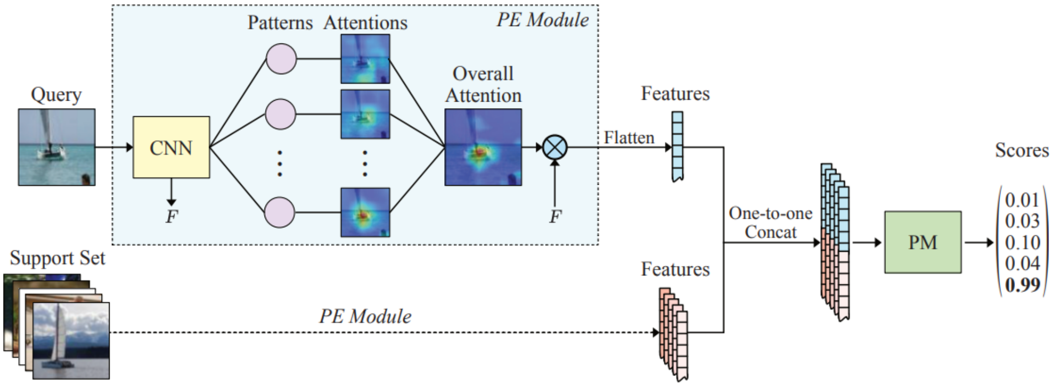
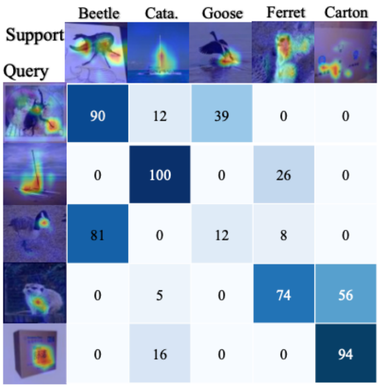


Figure 7. A new few-shot learning method.



Support / Query	Beetle	Cata.	Goose	Ferret	Carton
	90	12	39	0	0
	0	100	0	26	0
	81	0	12	8	0
	0	5	0	74	56
	0	16	0	0	94

Figure 8. FSL performance.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Wang Bowen、Li Liangzhi、Nakashima Yuta、Kawasaki Ryo、Nagahara Hajime、Yagi Yasushi	４．巻 9
２．論文標題 Noisy-LSTM: Improving Temporal Awareness for Video Semantic Segmentation	５．発行年 2021年
３．雑誌名 IEEE Access	６．最初と最後の頁 46810～46820
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3067928	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Wang Bowen、Li Liangzhi、Manisha Verma、Nakashima Yuta、Kawasaki Ryo、Nagahara Hajime
２．発表標題 MTUNet: Few-shot Image Classification with Visual Explanations
３．学会等名 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Responsible Computer Vision Workshop（国際学会）
４．発表年 2021年

１．発表者名 Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, Hajime Nagahara
２．発表標題 SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition
３．学会等名 IEEE/CVF International Conference on Computer Vision (ICCV)（国際学会）（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Github Code for the Semantic Segmentation
https://github.com/wbw520/NoisyLSTM
Github Code for the Few-shot Learning
https://github.com/wbw520/MTUNet

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関