摘要

每年全球范围内进行的手术的数量多达数亿，而医生的手术技能是影响病人术后康复状况的关键因素之一。为了保证患者福祉和减少临床失误，手术技能评价是外科医生培训和考核等环节中不可或缺的组成部分。传统的手术技能评价依赖于高年资医生的人工评价，包括直接观察和使用结构性打分体系等方式，但这种人工评价具有效率低和可重复性差的局限性，因此更加高效和可重复的自动化评价愈发地受到研究者们的关注。同时，近年来腹腔镜手术和机器人辅助手术的广泛应用积累了大量的手术视频数据，也为手术技能的自动化评价创造了良好的契机，本论文针对基于视频的手术技能自动化评价开展研究。

手术技能具有复杂的内涵，从定义出发可以将其拆解为操作技能、过程技能、间接指标三个技能方面。因此，本论文从宏观上确立了手术器械运动、手术事件分布、手术视野质量三个与上述技能方面相对应的评价方面，并设计了手术器械分割算法、手术事件识别算法、手术视野评价算法以获取各个评价方面的特征，最终形成了一个多特征融合的手术技能评价框架。本文具体的研究内容如下：

手术数据集构建：通过与北京大学肿瘤医院合作，本论文构建了两个来自临床真实场景的具有详细标注的手术数据集。其一为胃癌手术幽门下区数据集，包括 57 个胃癌手术片段的视频，标注涵盖了手术视野质量以及多项手术技能打分；其二为胃癌手术全过程数据集，包括 20 个完整胃癌手术的视频，标注涵盖了手术视野质量、手术技能打分以及各类手术事件。上述两个临床数据集为本论文的研究奠定了基础。

手术器械分割：为了从器械运动评价手术技能，本文提出了基于锚点生成和语义扩散的无监督器械分割方法，该方法首先以多种简单的视觉线索为基础生成锚点作为伪标签，然后利用视频帧间的时序语义相关性进一步地增强监督信号，从而实现了无监督的器械分割效果。实验表明，在不使用任何人工标注进行训练的前提下，本算法在 EndoVis-RobInstrument 数据集上可以取得 0.71 IoU 的优异性能。

手术事件识别：为了从事件分布评价手术技能，本文提出了使用强化学习的事件识别方法和利用多分支网络的弱监督事件识别方法。前者将手术事件识别任务建模为一个连续决策过程，并将时间连续性集成到了强化学习的动作集合与奖励函数的设计中，从而在实验中减少了预测结果的不连续现象；后者则设计了多分支神经网络结构和难负样本挖掘策略以应对弱监督事件识别中的完整性建模和上下文区分这两大难点，实验表明，该算法在 THUMOS-14 数据集和 ActivityNet 数据集上分别取得了 11% 和 28% 的性能提升。

手术视野评价：为了从手术视野评价手术技能，本文首先分析发现手术视野质量可以作为手术技能的良好间接指标，其具有与总体技能的高相关性以及评价者间的高一致性。然后本文提出了基于手术视野质量评价手术技能的算法，在训练该算法时本文同时使用了有监督的回归损失函数以及无监督的排序损失函数。实验表明，该算法在胃癌手术幽门下区数据集上的预测结果与医生打分间的相关性可达 0.595，这一性能甚至优于低年资医生的人类水平。

多特征融合的手术技能评价：本文依托上述内容进一步地提出了多特征融合的手术技能评价框架，该框架以上述各个评价方面得到的特征序列作为输入，并通过多路径神经网络对其进行前融合与后融合，并建模不同评价方面之间的依赖关系，从而对手术技能做出精准评价。除此之外，本框架还利用了自监督的视频预测编码机制以降低算法对训练数据量的需求。在胃癌手术全过程数据集上的实验表明，多特征融合的手术技能评价效果明显优于单特征的手术技能评价效果。

综上所述，本文以临床数据集为基础，从手术器械运动、手术事件分布、手术视野质量三个评价方面共同入手，提出了一个基于视频的手术技能多特征评价框架，为合理全面评价高度复杂的手术技能贡献了新的思路。同时，本文在算法设计上着重采用了无监督、弱监督、自监督等方式，以帮助手术技能评价算法在临床标注数据规模较小的约束下取得更好的效果。

关键词：手术技能评价，手术视频分析，手术器械分割，手术事件识别，手术视野质量

Video-Based Surgical Skill Assessment

Daochang Liu (Computer Application Technology)
Directed by Prof. Yizhou Wang and Prof. Tingting Jiang

ABSTRACT

Hundreds of millions of surgeries are performed worldwide annually. The proficiency of the operating surgeon is a key factor affecting outcomes after surgery. To ensure patients' well-being and reduce clinical errors, surgical skill assessment has become an indispensable part in surgical training and credentialing. Conventional surgical skill assessment is undertaken manually by experts with direct observation or structured rating protocols. Such human assessment is slow and hardly reproducible. Therefore, automatic surgical skill assessment is drawing more attention from researchers for its better efficiency and repeatability. Meanwhile, the prevalence of laparoscopic and robot-assisted surgeries nowadays brings a large volume of surgery videos captured by the built-in cameras in surgical devices, which lay the foundation for automatic assessment approaches. This thesis works on automatic surgical skill assessment using surgical videos.

Surgical skills are complex and can be broken down into three aspects, i.e., technical skills, procedural skills, and skill proxies. Therefore, three assessment aspects are identified in line with the above skill aspects in this thesis, which are surgical instrument usage, intraoperative event pattern, and clearness of operation field. We then accordingly develop algorithms for surgical instrument segmentation, surgical event recognition, and surgical field assessment to computationally capture the feature of each aspect. Finally, a multi-feature unified framework is established to provide an accurate assessment of surgical skills. Details are as follows.

Surgical video datasets: In this thesis, with the help of Peking University Cancer Hospital, two surgical video datasets from real clinical scenarios with detailed annotations are constructed. The first dataset consists of 57 videos of the infrapyloric phase of gastrectomy surgeries, annotated with the clearness of operation field and surgical skill ratings. The second dataset includes 20 videos of the whole process of gastrectomy surgeries, with annotations of the clearness of operation field, surgical skill ratings as well as diverse surgical events.

Surgical instrument segmentation: To assess surgical skills by surgical instrument

usage, this thesis proposes an unsupervised instrument segmentation method via anchor generation and semantic diffusion. This method first generates anchors as pseudo labels using multiple coarse visual cues and then exploits inter-frame semantic correlation to further enhance the supervision signal. Experiments show that the proposed method achieves an excellent performance of 0.71 IoU on the EndoVis-RobInstrument dataset without using a single manual annotation for training.

Surgical event recognition: To assess surgical skills by intraoperative event pattern, this thesis proposes a fully supervised event recognition method based on deep reinforcement learning and a weakly supervised event recognition method based on a multi-branch neural network. The former formulates the event recognition task as a sequential decision-making process. Temporal consistency is integrated into the action space and reward function of reinforcement learning to reduce over-segmentation errors. The latter explicitly tackles two major challenges in weakly-supervised event recognition, i.e., completeness modeling and context separation, by a multi-branch neural network and a hard negative mining strategy. Experiments show that the algorithm obtains 11% and 28% performance gains compared to prior works on THUMOS-14 and ActivityNet datasets respectively.

Surgical field assessment: To assess surgical skills by the clearness of operation field, this thesis first identifies the clearness of operation field as a good skill proxy because of its high correlation with overall skills and high consistency between annotators. We then propose an algorithm for assessing surgical skills through this proxy, which is trained using both a supervised regression loss and an unsupervised rank loss. Experiments show that the algorithm achieves 0.595 Spearman correlation with the ground truth on our infrapyloric-phase dataset, which even exceeds the human performance of junior surgeons.

Multi-feature unified surgical skill assessment: On the basis of the above works, a multi-feature unified surgical skill assessment framework is proposed. The framework comprises multiple paths in parallel, with each corresponding to an assessment aspect. Feature sequences extracted by the above algorithms are taken as inputs for the paths and are aggregated by both early-fusion and late-fusion schemes. The dependency relationships among different aspects are specially modeled by a path dependency module. In addition, the framework also utilizes a self-supervised predictive coding mechanism to learn without annotations. Experiments show that such multi-feature assessment is better than single-feature assessment. The framework achieves 0.565 Spearman correlation on our whole-process dataset.

In summary, this thesis proposes a unified framework for surgical skill assessment which

combines three different aspects of surgical skills, contributing new ideas on handling the huge complexity of surgical skills. Meanwhile, this thesis also focuses on using unsupervised, weakly-supervised, and self-supervised techniques in our algorithms to obtain improved performance with limited labeled data.

KEYWORDS: Surgical Skill Assessment, Surgical Video Analysis, Surgical Instrument Segmentation, Surgical Event Recognition, Clearness of Operation Field