IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Survey On Automatic Surgical Phase Recognition And Tool Identification

[1]Siddeswary Yadav S T, [2]Deepthi Murthy T. S
[1]Research Scholar, [2]Associate Professor
[1]School of Electronics and Communication Engineering,
[1]Reva University, Bangalore, India

Abstract: Robotic-assisted surgery (RAS) has emerged as a transformative force in modern surgical practices, particularly in minimally invasive surgery (MIS). This survey paper explores the evolution, current state, and prospects of RAS, underscoring its role in augmenting surgical precision and overcoming the constraints of conventional MIS techniques. Despite these advantages, RAS is seen without its challenges. Significant difficulties such as high operational costs, limitations in haptic feedback, and potential latency issues between control interfaces and robotic mechanisms are critically analysed. In this comprehensive review, we identify and discuss key research gaps within the domain of RAS. These include the need for advanced feature extraction methods capable of capturing essential details in surgical procedures, improved temporal and spatial modelling techniques, and the development of more efficient computational strategies to enhance the practicality of RAS systems. Additionally, this paper explores the intersection of RAS with surgical phase recognition technologies, a critical component in refining surgical workflows and augmenting real-time decision-making, as well as the importance of deep learning methodologies in advancing surgical phase recognition, highlighting their potential to significantly elevate the accuracy and efficiency of RAS.

Index Terms - Robotic-assisted surgery (RAS), Minimally invasive surgery (MIS), Temporal and spatial modeling, Surgical phase recognition technologies, Deep Learning.

I. INTRODUCTION

"Robotic-assisted surgery" refers to a surgical technique in which robotic technology is utilized to perform treatments. This surgical sub-specialty was framed to improve a surgeon's procedural skills and address the drawbacks of minimally invasive surgery (MIS). Even though it's frequently linked to minimally invasive surgery (MIS), open surgery occasionally makes use of robotic minimally invasive surgery (RMIS). Because robotic surgery offers more benefits than traditional Management Information Systems (MIS), it is considered more advantageous. The advantages that patients receive from minimally invasive surgery (MIS) are comparable. Following their hospital stays, patients should recover more quickly and also track the decrease in discomfort, tension, and scarring. Additionally, there will be a significant decrease in the risk of bleeding and infection. Surgeons benefit greatly from the increased visual capabilities, increased accuracy, and increased dexterity that the Robotic Minimally Invasive Surgery (RMIS) system offers. The increased agility that a robotic surgical system offers the surgeon is one of its main benefits. The use of robotic arms makes it possible to do jobs in constrained areas [1-2].

Furthermore, the robot doesn't get tired, which means the surgeon may work longer and in a more ergonomically comfortable posture during surgeries. The capacity to execute exact motions during an operation—a capability that traditional surgery cannot match—is one of the benefits of robotic surgery [3]. The surgeons' control over the robot's arms, as opposed to their own, is responsible for the increased precision and mobility in the small area. Furthermore, since robots never get tired, the accuracy of their arms is unaffected. The presence of supporting components like sensors and haptic feedback guarantees the precision of the robot arms' motions [4]. The increased visibility that robotic surgery provides is an additional benefit.

This is made possible by the robotic equipment's incorporation of cameras, which record and provide high-quality pictures of the patient's surgical operation. Moreover, the camera may be adjusted, providing the surgeon with a new perspective by allowing it to be repositioned. This gives the surgeon a better perspective of the process as it is being performed. Even though RMIS has many benefits, a few obstacles are preventing its wider implementation.

The haptic feedback restrictions and related expenses are among the robotic system's shortcomings. Because they are more expensive to operate than other systems, robotic systems are used less frequently. A surgeon's ability to maintain optimal force control may be hampered by inaccurate haptic feedback, which might lead to difficulties throughout the surgical operation. Furthermore, there's a chance that the robot and computer will experience lag. It is necessary to evaluate the existing constraints of robotic systems in great detail before they are standardized and approved for general use. To address any lingering issues, it is also crucial to have in-depth conversations and carry out more tests [5]. Here, robotic surgical systems are mentioned.

- **Zeus:** One of the top suppliers of medical equipment and technology is Zeus Surgical Equipment. Zeus Surgical Equipment provides a broad selection of surgical tools and places a high priority on innovation and quality. A robotic tool made especially for endoscopes is called the Automated Endoscopic Device for Optimal Positioning (AESOP). The FDA-authorized robot-assisted surgical system in question was the ground-breaking apparatus in dispute. The year 1994 saw Computer Motion Inc. release it onto the market. Figure 1 shows the Zeus Surgical System.

Figure 1 Zeus Surgical System

- **Da Vinci Surgical System:** The first robotic surgical system for general laparoscopic surgery, the Da Vinci system from Intuitive Surgical Inc., has received FDA approval. Subsequently, a greater variety of surgical procedures, such as thoracic, head & neck, colorectal, urologic, and cardiac surgeries, have been performed using this technique. Figure 2 shows the Da Vinci Surgical System.

Figure 2 Da Vinci Surgical System

- **Raven:** Figure 3 shows the beginning development of RAVEN at the University of Washington in 2002. The goal of the Raven system's design was to maximize the transfer of forces and position. Utilizing a large database of position and force data from laparoscopic surgeries, this was accomplished.

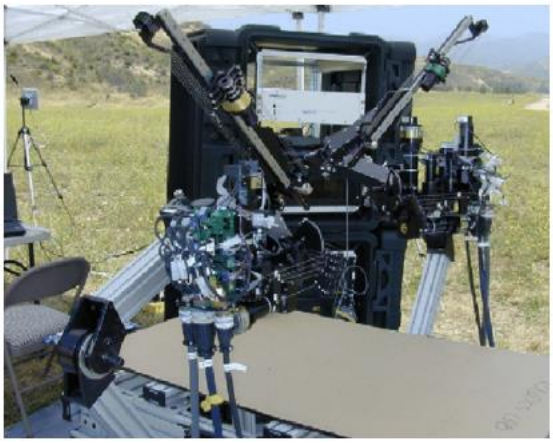

Figure 3 Raven

- **Flex Robotic:** One unique robotic arm made by Medrobotics Corporation is the Flex Robotic System, as seen in Figure 4. The Flex Colorectal Drive, a component of the Flex Robotic Colorectal System, may be used to move it along a non-linear route.

Figure 4 Flex Robotic

- **Sport:** Single-port laparoscopic surgery may be performed with the robotic surgical system SPORT. The tool's removable end effector tips provide it with a great degree of versatility. The surgeon may set up and arrange the workstation, and a flat-screen display will provide visual feedback. Figure 5 shows the Sport.

Figure 5 Sport

The significance of surgical phase identification for assessing and optimizing surgical workflow has made it a key topic in the field of computer-assisted interventions (CAI). Technology for real-time surgical phase identification is critical to the creation of context-aware systems. These devices can automatically update surgeons on the status of their procedures and send out alerts if there are any anomalies in the surgical process. Furthermore, context-aware systems are essential to improving human experiences. Videos are made up of a sequence of single frames, or still images, that are played backward and forward over a predetermined amount of time. The pictures that are displayed provide light on the ideas of time and place as well as the characteristics of human-object interactions. In a surgical video, the physician plays the part of the subject who alters the item, which stands for the operating field, to accomplish a certain goal [6]. Because surgical operations are performed regularly around the world, a significant amount of surgical video footage and related metadata have been collected. Advanced criteria for data management and organization are necessary due to the growing use of surgical video data applications. Furthermore, as surgical video data finds more and more uses, there is a growing demand for efficient data governance and the deployment of technologies like computer vision, machine learning, and artificial intelligence (AI) to facilitate deeper data analysis [7].

Physicians can get support from the Computer-Assisted Surgery (CAS) system during the intra-operative and post-operative phases. It does this by automatically determining the tool and surgical phase. By identifying rare occurrences and different variants, intra-operative identification can provide doctors with real-time caution. By facilitating effective communication among surgical team members, the system can assist less experienced surgeons in their decision-making [8]. Online recognition can improve OR resource management. Understanding the present surgical workflow and the particular instrument being used is required to get an estimate of how long the surgery will take to complete. Because of this feature, operating room efficiency is increased and patient wait times are decreased as clinical staff may proactively prepare for the next patient. Furthermore, post-operative recognition might improve the productivity of labor-intensive manual jobs that take a lot of time and effort to complete, such as indexing video databases and writing surgical reports. An indexed record of surgical procedures may improve the surgeon's training, review, and competency assessment. To enhance the surgical procedure, statistical information may be derived from the completely annotated database [9]. There are several obstacles in the way of developing automated methods to precisely identify the surgical phase and identify the presence of tools from the surgical film. A wide range of surgical tools is available, covering several special cases, including partial appearances and tool overlap. Considerable surgical treatments result in minor differences between phases and considerable oscillations within a specific phase.

Tool action and gas generation can cause the surgical sights to become partially or completely covered, especially if the camera lens is smeared with blood. The identification responsibilities are further complicated by the inclusion of noise and artifacts during the subsequent lens-cleaning procedure. To overcome the aforementioned problems, researchers used a variety of manual procedures in earlier experiments. These methods included the use of intensity values, combinative descriptors, and gradient magnitude [10]. However, the empirical creation of these low-level characteristics is not able to properly capture the fine distinctions seen in surgical recordings, since it significantly depends on domain experience. Since the emergence of deep learning, several attempts have been made to adapt convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for the aim of surgical video analysis [11].

The majority of deep learning methods now in use treat phase and tool identification tasks as discrete entities, ignoring their intrinsic connection. Using the appropriate tool sets at the appropriate phases of surgery is crucial for surgeons to comply with the regulations controlling the surgical process. There is a strong correlation between the surgical phase and tool use. The dissection processes used in the cholecystectomy approach [12] usually require the use of hooks. On the other hand, clippers and scissors are used throughout the cutting and clipping stages. Prior research that directly used binary instrument usage signals has successfully proved the usefulness of tool information in phase recognition [13]. Tool and phase recognition are both included in the multi-task framework that was created. During the feature learning phase, this architecture incorporates tool information and shares early layers. The favorable outcome implies that better phase identification and tool presence recognition depend on making the most of it [14]. The interdependence of many work functions frequently involves a high degree of complexity. In the setting of surgical films, it is typical for one tool to be used more than once, and different instrument combinations may be used at different stages of an operation. In light of the situation, it's critical to recognize that the previously described methods have significant shortcomings that could make it more difficult for them to accurately capture the relationship. [15] Present a method that applies temporal restrictions on phase prediction using a hidden Markov model

(HMM) in the context of video-based activities. This method is not the same as adding sequential data while the network is being trained.

The rapid evolution of robotic-assisted surgery (RAS) in the field of minimally invasive surgery (MIS) serves as a motivation for the comprehensive survey. This advancement is not just a technological leap but a paradigm shift in surgical practices, offering enhanced precision, reduced patient trauma, and improved postoperative outcomes. However, the integration of RAS into mainstream healthcare is hindered by significant challenges, including high costs, technical limitations, and the need for advanced computational strategies. Our motivation is driven by the potential to overcome these barriers, leveraging pioneering research in feature extraction, spatial-temporal modeling, and deep learning. Inspired by the prospect of refining surgical workflows and decision-making processes, ultimately leading to more effective, efficient, and patient-centric surgical care. This survey aims to not only outline the current state of RAS but also to illuminate the path forward, identifying key areas for future research and development of robotic systems in surgery. Further contribution is given as follows.

- It discusses the evolution, current state, and prospects of RAS, highlighting its benefits and challenges.
- The survey identifies key research gaps in RAS, such as the need for advanced feature extraction methods, improved temporal and spatial modeling techniques, and more efficient computational strategies.
- It also emphasizes the importance of adaptive approaches for the dynamic nature of surgical environments and explores the integration of RAS with surgical phase recognition technologies, underlining the role of deep learning in advancing these areas.
- The conclusion underscores the significance of addressing these research gaps for the future advancement of RAS, aiming to enhance surgical workflows and decision-making processes.

2. RELATED WORK

In the initial stages of studying surgical phase recognition from surgical movies, the focus was on utilizing hand-crafted features. These features encompassed pixel values and intensity gradients, spatial-temporal features, as well as features that were composed of color, texture, and form. Previous studies have employed different linear statistical models to capture the temporal information in surgical recordings. These models include Conditional Random Fields, hierarchical HMM, left-right HMM, Hidden semi-Markov Model, and Dynamic Time Warping [16]. The performance of these systems, however, is constrained by the low-level characteristics that have been established using empirical methods. In recent years, there has been notable advancement in the capability of neural networks to extract spatial and temporal data from surgical videos. This technological advancement has facilitated the identification and categorization of different phases within surgical procedures. The techniques can be classified into two distinct groups. The primary goal of this category is to effectively model both temporal and spatial characteristics by employing frame-wise labeling techniques. In their study, they utilized ResNet for feature extraction at the video level. Their findings showcased the effectiveness of this approach in accurately detecting surgical phases. In their study, [17] presented SV-RCNet, a comprehensive framework designed to progressively train spatial-temporal characteristics for the specific task of surgical phase identification. The system incorporates the integration of ResNet and an LSTM module. The researchers have developed TMRNet, a memory bank that has been specifically designed to integrate long-range and multi-scale temporal features. The main objective of this development is to accurately identify surgical phases.

The aim of [18] was to accurately capture the prolonged temporal dynamics linked to surgical procedures. In the study conducted by [18], a hybrid embedding aggregation transformer is employed to augment the significance of spatial characteristics in phase identification, while simultaneously capturing temporal information. In the domain of surgical phase recognition, several studies have utilized a multistage design approach. This methodology involves the integration of a refinement stage following the initial predictor phase. The objective of the refinement stage is to accurately correct any misclassifications that might have taken place during the predictor stage. The multistage temporal convolution network (MS-TCN) was adapted for online surgical scenarios by TeCNO through the use of dilated and causal convolutions. In their study, [19] put forward an alternative training approach. They found that using MS-TCN directly did not result in significant enhancements in performance. Another category employs supplementary data, such as the implementation of multitasking learning techniques, to enhance the performance of surgical phase recognition.

In their study, [20] presented the application of multitasking in the execution of a shared task. The methodology employed in this study encompassed phase recognition and tool presence detection to facilitate feature extraction. In their forecasting model, they employed a ResNet to make predictions on binary outcomes related to the presence of tools. The predictions and characteristics were subsequently combined to enhance the process of phase recognition. The MTRCNet-CL utilizes a distinct correlation loss to explicitly articulate the connections between tool presence and phase categorization. The study conducted by [21] involved the integration of multiple cues, such as management tools, ontology, and production norms, along with tool information, to enhance performance optimization. Multiple studies have been conducted to perform supplementary analysis for extracting optical fluxes and integrating motion data. This is done to improve the learning capabilities of the model. The implementation of these techniques leads to increased costs for multitasking annotations or introduces computational intricacies when incorporating new modalities, such as optical flows.

Earlier approaches have developed fixed multi-scale sliding windows, which are commonly used as recommendations for video grounding or temporal action localization. A recent study was conducted by researchers to construct an input-level frame pyramid. This was achieved by sampling frames at various temporal speeds. Furthermore, distinct networks were employed to extract frames from individual levels of the pyramid, facilitating the capture of pertinent mid-level features. The final prediction was generated by combining these features [22]. The implementation of these techniques necessitated the integration of supplementary networks, which could lead to increased computational costs. In their study, they have utilized a singular input to effectively capture visual tempos across various feature levels. This approach was inspired by the feature pyramid network (FPN). To adequately handle various temporal scales, the researchers utilized a feature pyramid network that integrates the downsampling of features over time. Table 1 shows the survey table for surgical phase identification.

Table 1 Survey table for surgical phase identification

Reference	Methodology	Advantage	Disadvantage	Research Gap
[11], [12]	Hand-crafted Features	Simple to implement	Limited to basic image properties	Need for more complex feature extraction
[13]	Left-Right HMM	Good at sequence analysis	Limited in capturing complex temporal patterns	Requires advanced temporal modeling techniques
[14]	Hidden Semi-Markov Model	Better temporal dynamics	May not fully capture intricate spatial details	Integration of spatial information
[3], [15]-[17]	Conditional Random Fields, Dynamic Time Warping	Effective in spatial-temporal relation modeling	Complex to implement and tune	Need for more efficient computational models
[25]-[27]	Multi-scale Sliding Windows	Useful for action localization	Fixed temporal window limits flexibility	Development of adaptive temporal analysis methods

The authors [23] propose a solution for the temporal action segmentation problem. They introduce a multi-stage temporal convolution network (MSTCN) that utilizes cascaded dilated 1D convolutions to capture long-range temporal information by expanding the receptive fields. This feature enables the collection of long-term

data on movies. They have proposed a method for long-range temporal order verification that allows for the isolation of activities from their context in a self-supervised manner. This approach effectively reduces the need for costly manual annotation in the analysis of long movies. Existing models can be categorized into two distinct categories. The initial category comprises single-stage models that utilize input visual information to generate prediction outcomes. Several studies have utilized various techniques such as conditional random fields, dynamic temporal warping, and different variants of Hidden Markov Models (HMM) to analyze retrieved visual features. The RNN model follows an end-to-end approach. It begins by utilizing a highly complex ResNet to extract visual attributes from each frame. Subsequently, the model employs an LSTM network to capture the temporal dependencies between consecutive frames. The second category comprises multi-stage models that employ an additional refinement step on top of the prediction findings to further enhance their performance. [24] Introduced a multi-stage architecture for the problem of surgical phase recognition. A subsequent causal Temporal Convolutional Network (TCN) is utilized to enhance the accuracy of predictions. This is done following the initial use of a causal TCN to generate preliminary predictions based on pre-extracted Convolutional Neural Network (CNN) features. The multi-stage design is considered to be well-suited for addressing the challenge of surgical phase identification. Multi-stage architecture networks have been widely employed in various computer vision applications that involve complex patterns. These applications include action segmentation and human posture estimation. In certain cases, the initial predictions generated by the predictor stage may exhibit errors that deviate from the inherent patterns in the data. This can occur due to the presence of visually complex characteristics that are challenging to identify.

For instance, instances of minuscule over-segmentation errors may occur during continuous motion, or there may be deviations in the human posture estimate findings that do not align with the connections between joints. The refining stage plays a crucial role in enhancing the accuracy of the initial predictions φyp [25]. In the refining step, only the initial predictions are utilized as input. This is done to avoid any interference from noisy visual characteristics. By focusing solely on the fundamental patterns in the data, the refinement process can be more effective. Furthermore, it has been observed that surgical video materials exhibit a wealth of temporal patterns and organization. The utilization of the intricate temporal patterns to enhance predictive capabilities has served as a source of inspiration for various endeavors. They have developed a mapping model that can be used to determine the phase label of the hard frames that have been identified, as previously expected. The accomplishments of the researchers serve as evidence that the implementation of a multi-stage design can rectify misclassifications that arise due to ambiguous visual indications during the predictor step [26]. Table 2 shows the survey table for surgical phase classification.

Table 2 Survey table for surgical phase classification

Reference	Methodology	Advantage	Disadvantage	Research Gap
[3], [18]	ResNet	Effective at video-level feature extraction	May overlook finer details in large datasets	Need for more detailed feature extraction
[19]	SV-RCNet	Combines spatial and temporal learning	Complexity in model training and tuning	Balancing model complexity with performance
[5]	TMRNet	Captures long-range temporal dynamics	Potential for overlooking short-term variations	Inclusion of short-term feature analysis
[6]	Hybrid Embedding Aggregation Transformer	Enhances spatial feature recognition	May be less effective in complex scenarios	Improving performance in diverse surgical contexts
[7], [9], [20]	MS-TCN	Adapted for real-	Initial misclassifications	Refinement of

		time scenarios	in the predictor stage	prediction accuracy

Surgical process modeling (SPM) has various benefits owing to its ability to identify separate surgical phases. Furthermore, the possibilities of SPM will expand with the advanced image recognition of deep learning [27]. Recognition technology for surgical phases using deep learning has been used in a variety of cases; for instance, predicting an operation's end time with an image of the surgical field, supporting surgeons' intraoperative decision-making, indexing surgical videos in a database, and assessing operation skills using videos. Notably, a deep learning model used in an operating room must have high versatility for an unknown image. Recently, deep learning systems to assist surgeons in decision-making have undergone remarkable developments, and the demand for surgical phase recognition techniques will increase shortly.

Similarly, related studies [28] have reported that surgical tools provide effective information to improve the recognition accuracy of the surgical phase. Importantly, in this method, using surgical tools to identify the surgical phase, the recognition accuracy often declines owing to the different colors of the hook shaft of endoscopic instruments. Additionally, blood on surgical tools and manipulations behind the organs are factors decreasing the recognition accuracy of the surgical phase. Furthermore, after upgrading the appearance of a surgical tool, the author must reconstruct the learning model by repeating a series of development cycles, such as annotation, training, and evaluation. If the learning model has already been embedded in a commercially available medical device, the reconstructed learning model must undergo regulatory examination at each redesign related to the appearance of a surgical tool. Considering the cost involved in updating the learning model, accurately recognizing the surgical phase without relying on the information provided by surgical tools is important to predict the surgical phase. EndoNet achieved approximately 0.82 overall accuracy for surgical phase recognition in laparoscopic cholecystectomy (LC), in which the features of an image from an endoscopic camera and a surgical tool are used to predict the surgical phase. The authors used the open datasets Cholec80 and EndoVis, which contain the video data of LC performed in a single facility. Additionally, the authors adopted long short-term memory (LSTM) in a recurrent neural network to estimate the surgical phase while considering the surgical phase to a certain point, resulting in 0.963 recognition accuracy. Also, the authors proposed a deep learning model with LSTM to estimate the remaining surgery duration intraoperatively [29]. However, it is considered that LSTM is not desirable to intraoperatively identify the surgical phase because unexpected intraoperative events happen frequently. In this regard, no redundant phase between the surgical phases was defined in either Cholec80 or EndoVis [30]; therefore, the benefits of LSTM were limited in these datasets. With the development of the latest deep learning models, it has become possible to recognize the surgical process with high accuracy without using the recognition information of surgical instruments for decision-making. However, for extracorporeal images, misty images during sectioning, and out-of-focus images, it is difficult for even deep learning models to accurately estimate the surgical process from the information from a single image. Therefore, in addition to improving the accuracy of the learning model, postprocessing to estimate the surgical process is important for the clinical use of the learning model. In this study, we aimed to construct a deep CNN model that intraoperatively identifies the surgical phase in LC and can be available as embedded software in a medical device. To accomplish our purpose, the surgical phase was recognized using only the endoscopic images obtained in LC [17].

Recently, the use of a deep CNN model to reduce the incidence of BDI has been reported. [21] proposed an AI system that intraoperatively indicates the anatomical landmarks during confirming Calot's triangle; proposed an automatic assessment tool for CVS during dissection of Calot's triangle; developed a deep learning model that visually identifies safe (Go) and dangerous (No-Go) zones for liver, gallbladder, and hepatocytic triangle dissection during LC. The purpose of these applications is limited to the specified surgical phase of confirming Calot's triangle and Calot's triangle dissection. Authors assume the surgical phase recognition model would be expected as a trigger for these applications [24].

2.1. RESEARCH GAP

- **Improving Low-Level Feature Extraction**: Current methods relying on hand-crafted features like pixel values and intensity gradients are limited by their simplistic nature. There's a gap in developing more complex, automated feature extraction techniques that can capture intricate details in surgical videos.
- **Enhanced Temporal Modeling**: Existing models like the Hidden Semi-Markov Model and left-right HMM are limited in capturing complex temporal dynamics. Advanced temporal modeling techniques are needed to better understand and represent the sequential nature of surgical procedures.
- **Spatial Feature Integration**: While some methods focus on temporal information, integrating spatial information into these models can provide a more holistic understanding of surgical videos.
- **Computational Efficiency**: Techniques such as Conditional Random Fields and Dynamic Time Warping are computationally intensive. There's a need for more efficient computational models that maintain accuracy while reducing processing time.
- **Adaptive Temporal Analysis Methods**: Fixed temporal window methods like multi-scale sliding windows lack flexibility. Research is needed to develop adaptive temporal analysis methods that can adjust to the varying nature of surgical videos.
- **Detailed Feature Extraction in Large Datasets**: Methods like ResNet, while effective at a general level, may overlook finer details, especially in large datasets. Research is required to enhance the granularity of feature extraction.
- **Balancing Model Complexity and Performance**: Frameworks like SV-RCNet, which combines spatial and temporal learning, face challenges in balancing model complexity with performance. Simplifying these models without sacrificing accuracy is a key research gap.
- **Short-term Feature Analysis in Long-range Models**: Models capturing long-range temporal dynamics sometimes overlook short-term variations. Incorporating short-term feature analysis into these models could provide a more comprehensive understanding.
- **Performance in Diverse Surgical Contexts**: Spatial feature enhancement methods may underperform in complex surgical scenarios. Research is needed to optimize these methods for diverse and unpredictable surgical environments.
- **Reducing Reliance on Additional Data Inputs**: Many multi-task learning models require additional data inputs, such as tool presence, which increase the annotation cost and computational load. Developing methods that reduce reliance on such inputs while maintaining accuracy is a crucial research area.

3. CONCLUSION

In conclusion, robotic-assisted surgery (RAS) stands at the forefront of innovation in minimally invasive surgery (MIS), offering significant enhancements in precision and efficiency over traditional techniques. However, the evolution of RAS also brings to light challenges such as high operational costs and technical limitations, which necessitate ongoing research and development. Addressing key research gaps, particularly in advanced feature extraction, temporal and spatial modeling, and computational efficiency, is crucial for the future advancement of RAS. The integration of RAS with surgical phase recognition, strengthened by deep learning methodologies, presents a promising avenue for refining surgical workflows and improving decision-making processes. As we continue to navigate these challenges and explore these integrations, RAS is poised to redefine surgical standards, promising a future where surgical interventions are more accurate, efficient, and adaptable to the dynamic nature of clinical environments.

4. ACKNOWLEDGEMENT

I would like to express our sincere gratitude to all those who have supported and contributed to this research project. Primarily, I extend our heartfelt thanks to our guide for his unwavering guidance, invaluable insights, and encouragement throughout the research process. No funding is raised for this research.

REFERENCES

[1] Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, PA. (2021). Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer. In: de Bruijne, M., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science(), vol 12904. Springer, Cham. https://doi.org/10.1007/978-3-030-87202-1_57

[2] Czempiel, T. et al. (2020). TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks. In: Martel, A.L., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science(), vol 12263. Springer, Cham. https://doi.org/10.1007/978-3-030-59716-0_33

[3] Yi, F., Yang, Y., Jiang, T. (2023). Not End-to-End: Explore Multi-Stage Architecture for Online Surgical Phase Recognition. In: Wang, L., Gall, J., Chin, TJ., Sato, I., Chellappa, R. (eds) Computer Vision – ACCV 2022. ACCV 2022. Lecture Notes in Computer Science, vol 13844. Springer, Cham. https://doi.org/10.1007/978-3-031-26316-3_25

[4] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3575–3584, doi.org/10.48550/arXiv.1903.01945

[5] Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., & Jannin, P. (2016). Automatic data-driven real-time segmentation and recognition of surgical workflow. International journal of computer assisted radiology and surgery, 11(6), 1081–1089. https://doi.org/10.1007/s11548-016-1371-x.

[6] Y. Jin et al., "SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network," in IEEE Transactions on Medical Imaging, vol. 37, no. 5, pp. 1114-1126, May 2018, doi: 10.1109/TMI.2017.2787657.

[7] G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks," 2018, arXiv:1805.08569, doi.org/10.48550/arXiv.1805.08569

[8] Nwoye, C.I., Mutter, D., Marescaux, J. et al. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. Int J CARS 14, 1059–1067 (2019). https://doi.org/10.1007/s11548-019-01958-6

[9] Meireles, O. R., Rosman, G., Altieri, M. S., Carin, L., Hager, G., Madani, A., Padoy, N., Pugh, C. M., Sylla, P., Ward, T. M., Hashimoto, D. A., & SAGES Video Annotation for AI Working Groups (2021). SAGES consensus recommendations on an annotation framework for surgical video. Surgical endoscopy, 35(9), 4918–4929. https://doi.org/10.1007/s00464-021-08578-9

[10] Huaulmé, A., Jannin, P., Reche, F., Faucheron, J. L., Moreau-Gaudry, A., & Voros, S. (2020). Offline identification of surgical deviations in laparoscopic rectopexy. Artificial intelligence in medicine, 104, 101837. https://doi.org/10.1016/j.artmed.2020.101837

[11] Y. Lin, H. Yao, Z. Li, G. Zheng, and X. Li, "Calibrating label distribution for class-imbalanced barely-supervised knee segmentation," 2022, arXiv:2205.03644, DOI:10.1007/978-3-031-16452-1_11.

[12] Y. Li, Y. Duan, Z. Kuang, Y. Chen, W. Zhang, and X. Li, "Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation," in Proc. AAAI, 2022, pp. 4556–4562, DOI:10.1609/aaai.v36i2.20034.

[13] Yi, F., Yang, Y., Jiang, T. (2023). Not End-to-End: Explore Multi-Stage Architecture for Online Surgical Phase Recognition. In: Wang, L., Gall, J., Chin, TJ., Sato, I., Chellappa, R. (eds) Computer Vision – ACCV 2022. ACCV 2022. Lecture Notes in Computer Science, vol 13844. Springer, Cham. https://doi.org/10.1007/978-3-031-26316-3_25

[14] Premakumari Pujar, Ashutosh Kumar, Vineet Kumar, "Efficient plant leaf detection through machine learning approach based on corn leaf image classification" IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 13, No. 1, March 2024, pp. 1139~1148, ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i1.pp1139-1148.

[15] Sreedhara, S.H., Kumar, V., Salma, S. (2023). Efficient Big Data Clustering Using Adhoc Fuzzy C Means and Auto-Encoder CNN. In: Smys, S., Kamel, K.A., Palanisamy, R. (eds) Inventive Computation and Information Technologies. Lecture Notes in Networks and Systems, vol 563. Springer, Singapore. https://doi.org/10.1007/978-981-19-7402-1_25

[16] Pan, X., Gao, X., Wang, H., Zhang, W., Mu, Y., & He, X. (2023). Temporal-based Swin Transformer network for workflow recognition of surgical video. International journal of computer assisted radiology and surgery, 18(1), 139–147. https://doi.org/10.1007/s11548-022-02785-y

[17] Zhang, B., Ghanem, A., Simes, A. et al. Surgical workflow recognition with 3DCNN for Sleeve Gastrectomy. Int J CARS 16, 2029–2036 (2021). https://doi.org/10.1007/s11548-021-02473-3

[18] Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, PA. (2021). Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer. In: de Bruijne, M., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science (), vol 12904. Springer, Cham. https://doi.org/10.1007/978-3-030-87202-1_57

[19] Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N. (2021). OperA: Attention-Regularized Transformers for Surgical Phase Recognition. In: de Bruijne, M., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science(), vol 12904. Springer, Cham. https://doi.org/10.1007/978-3-030-87202-1_58.

[20] Bhattarai, B., Subedi, R., Gaire, R. R., Vazquez, E., & Stoyanov, D. (2023). Histogram of Oriented Gradients meets deep learning: A novel multi-task deep network for 2D surgical image semantic segmentation. Medical image analysis, 85, 102747. https://doi.org/10.1016/j.media.2023.102747

[21] Zisimopoulos, O. et al. (2018). DeepPhase: Surgical Phase Recognition in Cataract Videos. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science(), vol 11073. Springer, Cham. https://doi.org/10.1007/978-3-030-00937-3_31

[22] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C.-W. Fu, and P.-A. Heng, “Multitask recurrent convolutional network with correlation loss for surgical video analysis,” Medical image analysis, vol. 59, p. 101572, 2020, doi.org/10.1016/j.media.2019.101572.

[23] Nakawala, H., Bianchi, R., Pescatori, L. E., De Cobelli, O., Ferrigno, G., & De Momi, E. (2019). "Deep-Onto" network for surgical workflow and context recognition. International journal of computer assisted radiology and surgery, 14(4), 685–696. https://doi.org/10.1007/s11548-018-1882-8.

[24] [24] D. Sarikaya, K. A. Guru, and J. J. Corso, “Joint surgical gesture and task classification with multi-task and multimodal learning,” arXiv preprint arXiv:1805.00721, 2018, doi.org/10.48550/arXiv.1805.00721.

[25] Z. Shou, D. Wang and S. -F. Chang, "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1049-1058, doi: 10.1109/CVPR.2016.119.

[26] Ding, X., Wang, N., Li, J., Gao, X. (2021). Weakly Supervised Temporal Action Localization with Segment-Level Labels. In: Ma, H., et al. Pattern Recognition and Computer Vision. PRCV 2021. Lecture Notes in Computer Science(), vol 13019. Springer, Cham. https://doi.org/10.1007/978-3-030-88004-0_4.

[27] G. Li, J. Li, N. Wang, X. Ding, Z. Li and X. Gao, "Multi-Hierarchical Category Supervision for Weakly-Supervised Temporal Action Localization," in IEEE Transactions on Image Processing, vol. 30, pp. 9332-9344, 2021, doi: 10.1109/TIP.2021.3124671.

[28] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal Activity Localization via Language Query," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 5277-5285, doi: 10.1109/ICCV.2017.563.

[29] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell and B. Russell, "Localizing Moments in Video with Natural Language," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 5804-5813, doi: 10.1109/ICCV.2017.618.

[30] X. Ding et al., "Support-Set Based Cross-Supervision for Video Grounding," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 11553-11562, doi: 10.1109/ICCV48922.2021.01137.