International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2283-2288
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0336

Deep Learning Model for Instrument Detection in Medical Surgeries and Avoiding Mistakes

Pratiksha Rasal[1], Snehal Rakshe[2], Siddhi Shelke[3], K.S. Bhagwat[4]
[1,2,3,4]Dept. of IT VPKBIET, Baramati, Pune, India.
Emails: *pratiksharasalb@gmail.com[1], rakshesnehal130@gmail.com[2], shelkesiddhi27@gmail.com[3], keshav.Bhagwat@vpkbiet.org [4]*

Abstract

In the fast-evolving healthcare sector, accurate detection and classification of medical tools are essential for enhancing surgical efficiency and patient safety. This paper presents a novel approach to automatic medical device detection using advanced computer vision and deep learning, specifically the YOLOv8 model. The system is trained on a dataset containing various instruments like scalpels, forceps, and scissors, with data preprocessing, augmentation, and transfer learning techniques applied to boost performance despite limited training data. Designed for real-time operation in surgical environments, the model is evaluated using metrics such as accuracy, precision, and recall to ensure reliable performance.

Keywords: *Deep Learning, Instrument Detection, Object Detection, Skill Assessment, Surgical Performance.*

1. Introduction

Medical equipment varies greatly in shape, size, and function, posing particular problems to detecting systems. traditional techniques of instrument recognition frequently rely on manual identification, which is time consuming and susceptible to human error. As surgical procedures become more sophisticated the need for efficient and dependable instrument detection has never been higher. Recent advances in computer vision and deep learning present interesting solutions to this dilemma. YOLOV3, a powerful class of deep learning algorithms, has shown remarkable performance in picture classification tests, making it perfect for identifying medical items. By training models on big datasets, these systems can learn to distinguish between distinct tools with great accuracy. This article proposes a comprehensive framework for the automated detection of medical devices using computer vision. Deep learning models, such as convolutional neural networks (CNNs), have demonstrated exceptional effectiveness in a variety of image recognition tasks, including those involving medical imaging. Their ability to acquire hierarchical feature representations makes them ideal for identifying and segmenting surgical tools in complex surgical scenarios [1].

2. Literature Survey

Kyle Lam et al. [1] (2022) proposed a deep learning-based method for instrument detection and assessment of operative skill in surgical videos They employed Mask R-CNN on a novel dataset of 2600 annotated images from laparoscopic gastric band insertion procedures to evaluate both instrument identification and surgical skill metrics. The study is limited by its narrow scope, focusing only on gastric band insertion, with tests conducted on just two videos and data gathered from only three patients raising concerns about the model's generalizability across diverse surgical scenarios. Amy Jin et al. [2] proposed a region-based convolutional neural network approach for tool detection and operative skill assessment in laparoscopic surgical videos. Their method introduces a new dataset to enhance analysis and demonstrates improved performance over existing techniques. However, the study is limited by the variability of surgical tasks and gestures, which may not accurately represent real-world surgical performance, potentially affecting the model's general applicability. Shubhangi Nema and Leena Vachhani [3] reviewed AIbased surgical instrument detection and tracking technologies focused on automating dataset labeling for minimally invasive surgical skill assessment. The paper highlights recent advancements, ongoing challenges, and potential future directions in the field. However, a key limitation is the inability to reliably quantify surgical skill levels, despite progress in automating assessments through AI-driven methods. Romina Pedrett and Pietro Mascagni [4] conducted a systematic review of AI models used for technical skill assessment in minimally invasive surgery. The study evaluates model performance, external validity, and generalizability, while also identifying key challenges in automating surgical skill evaluation. A major limitation noted is the frequent lack of external validation, standardization, and generalizability of these models, which restricts their practical use in diverse surgical environments. Bareum Choi et al. [5] explored the use of convolutional neural networks (CNNs) for surgical tool detection in laparoscopic robot-assisted surgeries, aiming to enhance automation in tool tracking. While the approach shows promise, it faces challenges with variations in tool appearance, occlusions, and real-time processing, which can hinder its effectiveness in complex surgical environments. J. J. Corso and K. A. Guru [6] proposed a deep neural network-based method for detecting and localizing robotic tools in robot-assisted surgery videos. Their approach emphasizes region proposal and detection techniques to enhance accuracy and automation in surgical tool tracking. However, the method faces challenges related to processing speed, appearance variations, occlusions, robustness, and generalizability, which can impact its performance in dynamic surgical environments [7].

3. Current Methodologies

3.1 Dataset Collection and Preprocessing

The process begins with endoscopic images, which are analyzed using a pre-existing, annotated dataset containing labeled medical instruments for training the Mask R-CNN. The YOLOv8 algorithm is then applied, treating object detection as a single regression problem rather than separating it into stages like region proposal and classification. YOLOv8 outputs the spatial boundaries of the instruments within the images, which are subsequently used for workflow analysis. This includes examining instrument movements and usage patterns to better understand surgical procedures.

3.2 Instrument detection

Instrument detection is the process of accurately identifying and classifying different surgical instruments used during medical procedures. This technology plays a crucial role in enhancing the efficiency and safety of surgeries by ensuring that all instruments are accounted for and used correctly. Accurate detection also aids in automating surgical workflows and supporting real-time decision-making in the operating room.

3.3 Ensemble learning: Ensemble learning improves surgical

instrument detection Ensemble learning enhances surgical instrument detection by combining the strengths of multiple machine learning models to achieve better accuracy and robustness. Instead of relying on a single model, ensemble methods aggregate predictions from several models, reducing errors and improving overall performance. This approach is particularly effective in complex tasks like surgical instrument detection, where precision is critical for patient safety and surgical efficiency.

4. Related Work

4.1 Image-Based Surgical Instrument Detection

Recent advances in deep learning have led to significant improvements in image-based surgical instrument detection. Convolutional Neural Networks (CNNs) such as VGG, ResNet, and Efficient Net are commonly used for detecting and classifying instruments in surgical scenes. VGG and ResNet architectures are known for their strong feature extraction capabilities, but they are often associated with high computational costs [8].

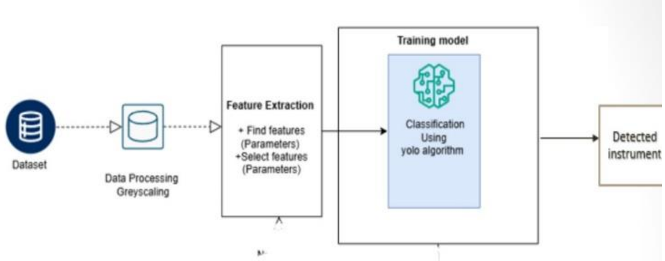

Figure 1 Architecture

EfficientNet addresses this limitation through compound scaling, which balances network depth, width, and resolution to achieve high accuracy with fewer parameters. This makes EfficientNet a suitable choice for real-time and resource-constrained surgical environments. Figure 1 shows Architecture.

4.2 Temporal Modeling for Instrument Tracking

In surgical procedures, tracking instruments across video frames is essential for ensuring continuity and safety. Traditional approaches have utilized Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)networks to capture temporal dependencies. While effective for short term tracking, these models often face challenges such as vanishing gradients, limiting their ability to model long-term sequences [9]. Transformer-based architectures, leveraging self-attention mechanisms, have emerged as a powerful alternative by capturing global temporal relationships without relying on sequential processing. These models offer improved accuracy in dynamic and complex surgical environments [11].

4.3 Ensemble Learning for Enhanced Detection

Ensemble learning has been applied to improve the robustness and accuracy of surgical instrument detection. By combining multiple models or classifiers, ensemble methods reduce the variance and improve generalization performance. Techniques such as bagging, boosting, and stacking allow integration of different model predictions, compensating for individual model weaknesses. This approach is particularly valuable in high-stakes medical applications, where precision and reliability are critical for avoiding errors during surgery.

5. Enhancements in Model Design

5.1 Advantages of EfficientNet

EfficientNet provides notable benefits when applied to surgical instrument detection tasks: **Efficient Scaling:** Its compound scaling method optimally balances depth, width, and resolution, yielding higher accuracy with fewer parameters critical for real time surgical applications [12]. **Transfer Learning:** EfficientNet models pre-trained on large datasets like ImageNet can be fine-tuned for medical instrument detection, enabling faster convergence and enhanced generalization on limited surgical data. **Geometric Feature Handling:** The architecture is capable of extracting robust spatial features of instruments through geometric transformations, improving localization and classification performance.

5.2 Advantages of Transformers

Transformers bring significant benefits for temporal modelling: **Self-Attention Mechanisms:** These mechanisms allow the model to selectively attend to relevant frames and tool states over time, improving decision-making in sequential frames. **Scalability:** Transformers can process large-scale surgical video datasets in parallel, making them ideal for high resolution, long-duration surgery recordings. **Non-Sequential Processing:** Unlike RNNs or GRUs, Transformers do not suffer from vanishing gradient issues, and can effectively capture long-range dependencies vital for tracking instruments across lengthy procedures [10].

5.3 Fusion Techniques

Model fusion enhances prediction reliability in multi-modal. surgical settings: **Comparison of Methods:** Techniques such as weighted fusion, stacking, concatenation, and soft voting enable integration of predictions from CNNs and temporal models. **Advantages of Weighted Fusion:** By dynamically adjusting model contributions based on confidence or past performance, weighted fusion produces more accurate and balanced instrument detection outcomes, minimizing the likelihood of surgical errors. Figure 2 shows Precision and Recall vs IoU Threshold.

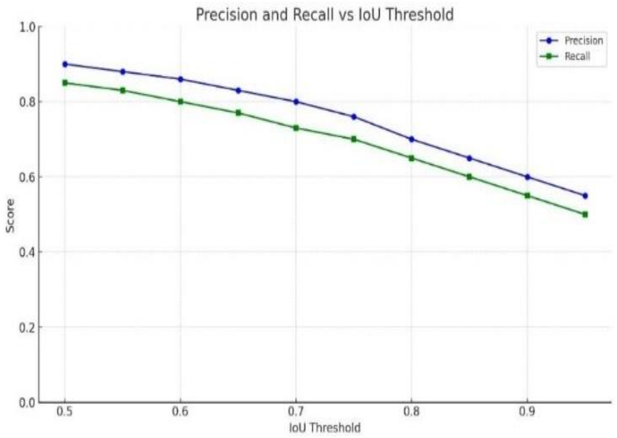

Figure 2 Precision and Recall vs IoU Threshold

6. Algorithms

6.1 YOLOv8

We employ YOLOv8, a state-of-the-art single-stage object detection algorithm, for real-time detection and localization of surgical instruments in endoscopic and laparoscopic imagery. YOLOv8 builds upon the efficiency of previous YOLO versions while introducing improvements in model architecture, anchor-free detection, and scale-aware training, making it well-suited for the challenges of the surgical domain, such as occlusions, tool overlap, and varying lighting conditions. We fine-tuned pretrained YOLOv8 models on annotated datasets comprising common surgical instruments using transfer learning, optimizing for mean Average Precision (mAP) and inference speed [14]. The model demonstrated robust performance in identifying multiple instruments with high precision and recall, even in complex surgical scenes. This approach enables accurate and real-time instrument detection, supporting downstream applications in robotic surgery, surgical workflow analysis, and intraoperative decision support. training deep learning models for instrument detection on diverse and comprehensive datasets significantly improves detection accuracy and generalization. A well-curated dataset that incorporates a wide range of surgical contexts, instrument variations, and temporal dynamics is essential for building robust and reliable systems capable of operating effectively in real-world clinical settings. Figure 3 shows AlexNet CNN Architecture.

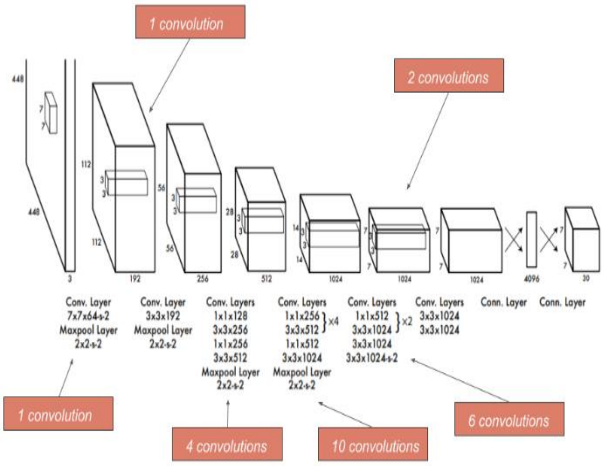
1 convolution
2 convolutions
448
7
7
448
3
112
112
192
56
56
256
28
28
512
14
14
1024
7
7
1024
1024
4096
30
Conv. Layer 7x7x64-s-2 Maxpool Layer 2x2-s-2
Conv. Layer 3x3x192 Maxpool Layer 2x2-s-2
Conv. Layers 1x1x128 3x3x256 1x1x256 3x3x512 Maxpool Layer 2x2-s-2
Conv. Layers 1x1x256 3x3x512 }x4 1x1x512 3x3x1024 Maxpool Layer 2x2-s-2
Conv. Layers 1x1x512 3x3x1024 }x2 3x3x1024 3x3x1024-s-2
Conv. Layers 3x3x1024 3x3x1024
Conn. Layer
Conn. Layer
1 convolution
4 convolutions
10 convolutions
6 convolutions

Figure 3 **AlexNet CNN Architecture**

7. Results and Discussion

7.1 Results

The developed endoscopic instrument detection system was evaluated on a dataset containing multiple surgical instruments commonly used in minimally invasive procedures. The model demonstrated the ability to detect and classify various tools such as graspers, scissors, clip appliers, needle holders, and Bulldog clamps across different surgical frames. Instrument detection confidence scores varying based on tool visibility, size, and dataset representation. Instruments with clear structure and frequent occurrence, such as graspers and scissors, showed high detection accuracy with confidence levels consistently above 0.80. In contrast, less frequently seen or partially occluded instruments, like the Bulldog clamp, exhibited lower detection confidence (e.g., 0.26), indicating a need for dataset balancing and further training. Despite these limitations, the overall performance suggests that the model is capable of multi-instrument recognition with reliable localization, forming a strong foundation for integration into surgical navigation, training analysis, or robotic assistance systems. Figure 4 shows Surgical Tool Detection – Bulldog Clip.

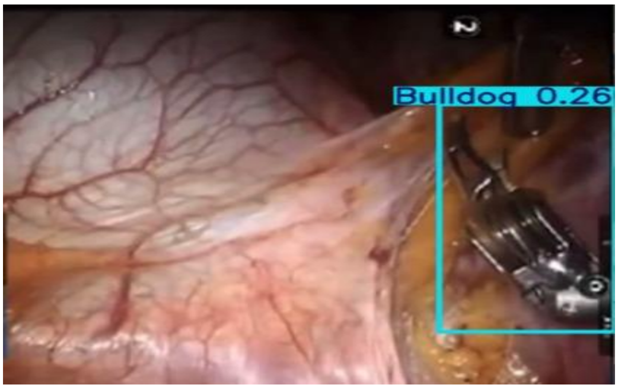

Figure 4 **Surgical Tool Detection – Bulldog Clip**

7.2 Discussion

The detection model performed well for commonly used instruments like graspers and scissors, achieving high accuracy due to their frequent appearance and distinct features [15]. However, detection accuracy decreased for less common or visually subtle tools like the Bulldog clamp, mainly due to limited training samples and challenging visual conditions such as occlusion and low contrast. These results highlight the need for a more balanced dataset, improved image quality handling, and possibly incorporating video-based context to enhance robustness. Overall, while the model shows strong potential, further optimization is needed for reliable use in real surgical settings.

Conclusion

Accurate and comprehensive detection of surgical instruments is critical for enhancing patient safety and preventing retained surgical items (RSIs). A robust detection system must be capable of identifying a diverse range of instruments, including forceps, scissors, scalpels, and sponges, while effectively handling variations in size, shape, and material. In this context, the YOLO (You Only Look Once) framework, optimized for real-time object detection, presents a highly suitable solution. Its ability to rapidly and accurately localize objects makes it ideal for monitoring surgical tools during and after procedures. Importantly, YOLO can also detect anomalies such as missing parts of instruments, enabling timely interventions and reducing the risk of complications. Leveraging such advanced deep learning models contributes significantly to surgical safety, supports operational decision making and minimizes the potential for human error in high stakes clinical environments.

Acknowledgements

We sincerely thank our guide, Mr. K.S.Bhagwat, for her invaluable guidance and continuous support throughout this project. Her insights and expertise were instrumental in the successful completion of our work. We extend our heartfelt gratitude to Mrs.S.A.Takle, Head of the Department of IT, and Dr. S. B. Lande, Principal of VPKBIET, Baramati, for providing us with the necessary resources and a conducive environment for research and development. Finally, we express our appreciation to the teaching staff of the Department of IT for their encouragement, constructive feedback, and unwavering support, which significantly contributed to the success of this work.

References

[1]. J. D. Birkmeyer et al.,"Surgical skill and complication rates after bariatric surgery" New England J. Med., vol. 369, no. 15, pp. 1434–1442, Oct. 2013.

[2]. N. Ahmidi et al.," Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," Int. J.Comput. Assist. Radiol. Surg., vol. 10, no. 6, pp. 981–991, Jun. 2015.

[3]. A. J. Hung et al.," Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes," J. Endourol., vol. 32, no. 5, pp. 438–444, May 2018.

[4]. A. Zia and I. Essa," Automated surgical skill assessment in RMIS training," Int. J. Comput. Assist. Radiol. Surg.," vol. 13, no. 5, pp. 731–739, May 2018.

[5]. R. C. King, L. Atallah, B. P. L. Lo, and G.-Z. Yang," Development of a wireless sensor glove for surgical skills assessment," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 5, pp. 673–679, Sep. 2009.

[6]. A. P. Twinanda, S. Shehata, "Szengel, and S. Zachow, "EndoNet: A deep architecture for recognition tasks on Laparoscopic videos," EEE Trans. Med. Imag., vol. 36, no. 1, pp. 86–97, Jan. 2017.

[7]. D. Sarikaya, J. J. Corso, and K. A. Guru, Detection and Localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection," IEEE Trans. Med. Imag., vol. 36, no.7, pp. 1542–1549, Jul. 2017.

[8]. B. Choi, K. Jo, S. Choi, and J. Choi," Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery,"in Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., Jul. 2017, pp. 1756–1759.

[9]. K. Mishra, R. Sathish, and D. Sheet," Learning latent temporal Connectionism of deep residual visual abstractions for

identifying surgical tools in laparoscopy procedures,"Proc. CVPR Workshops, 2017, pp. 2233–2240.

[10]. I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3Dconvolutional neural networks,," IEEE Access, vol. 12, 2024.

[11]. Shubhangi Nema,Leena Vachhani, "Surgical instrument detection and tracking technologies: Automating dataset labeling for surgical skill assessment" Front. Robot. AI, 04 November 2022.

[12]. M. Grammatikopoulou et al., """CaDIS: Cataract dataset for surgical RGB-image segmentation," in International Journal of Information Technology, vol. 14, no. 4,pp. 1631-1640, 2022.

[13]. N. Ahmidi et al. "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," Int. J.Comput. Assist. Radiol. Surg., vol. 10, no. 6, pp. 981–991, Jun. 2015.

[14]. J. D. Birkmeyer et al., """Surgical skill and complication rates after bariatric surgery," in IEEE Access, New England J. Med., vol. 369, no. 15, pp. 1434–1442, Oct. 2013

[15]. Amy Jin, "Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks,"arXiv:1802.08774v2 [cs.CV] , 22 Jul.2018