Original Article
Clinics in Orthopedic Surgery 2024;16:210-216 • https://doi.org/10.4055/cios23280

Assessment of Automated Identification of Phases in Videos of Total Hip Arthroplasty Using Deep Learning Techniques

Yang Jae Kang, PhD, Shin June Kim, MS*, Sung Hyo Seo, PhD†, Sangyeob Lee, DVM†, Hyeon Su Kim, BS*, Jun-Il Yoo, MD‡

Division of Bio and Medical Big Data Department (BK4 Program) and Life Science Department, Gyeongsang National University, Jinju,
**Biomedical Research Institute, Inha University Hospital, Incheon,*
†Biomedical Research Institute, Gyeongsang National University Hospital, Jinju,
‡Department of Orthopedic Surgery, Inha University Hospital, Inha University College of Medicine, Incheon, Korea

Background: As the population ages, the rates of hip diseases and fragility fractures are increasing, making total hip arthroplasty (THA) one of the best methods for treating elderly patients. With the increasing number of THA surgeries and diverse surgical methods, there is a need for standard evaluation protocols. This study aimed to use deep learning algorithms to classify THA videos and evaluate the accuracy of the labelling of these videos.

Methods: In our study, we manually annotated 7 phases in THA, including skin incision, broaching, exposure of acetabulum, acetabular reaming, acetabular cup positioning, femoral stem insertion, and skin closure. Within each phase, a second trained annotator marked the beginning and end of instrument usages, such as the skin blade, forceps, Bovie, suction device, suture material, retractor, rasp, femoral stem, acetabular reamer, head trial, and real head.

Results: In our study, we utilized YOLOv3 to collect 540 operating images of THA procedures and create a scene annotation model. The results of our study showed relatively high accuracy in the clear classification of surgical techniques such as skin incision and closure, broaching, acetabular reaming, and femoral stem insertion, with a mean average precision (mAP) of 0.75 or higher. Most of the equipment showed good accuracy of mAP 0.7 or higher, except for the suction device, suture material, and retractor.

Conclusions: Scene annotation for the instrument and phases in THA using deep learning techniques may provide potentially useful tools for subsequent documentation, assessment of skills, and feedback.

Keywords: *Arthroplasty, Hip, Deep learning, Surgical procedures*

As the population ages, there is a noticeable increase in the prevalence of hip diseases and fragility fractures. In light of this demographic shift, total hip arthroplasty (THA) has emerged as one of the most effective treatment options for elderly patients.[1,2] It is estimated that hundreds of thousands of THA procedures are performed each year in the United States alone.[3] The procedure is typically performed to alleviate pain and improve mobility in people with hip joint disorders such as osteoarthritis, rheumatoid arthritis, and avascular necrosis. Recently, fast recovery has been possible after THA with the development of bearing technology and the rapid development of minimally invasive technologies, including the direct anterior approach (DAA). In addition, the accuracy of surgery using navigation has been proven, it is widely used in the clinical field, and the surgical results of THA using robots have been reported.[4]

Received September 12, 2023; Revised October 23, 2023;
Accepted November 6, 2023
Correspondence to: Jun-Il Yoo, MD
Department of Orthopedic Surgery, Inha University Hospital, Inha University College of Medicine, 27 Inhang-ro, Jung-gu, Incheon 22332, Korea
Tel: +82-55-750-8688, Fax: +82-55-754-0477
E-mail: furim@hanmail.net
Yang Jae Kang and Shin June Kim contributed equally to this work.

To satisfy the needs of these patients, orthopedic resident education programs should train doctors to perform THA effectively, and it is necessary to establish a program development and evaluation system that can rapidly improve the learning curve. In the real clinical field, mastering THA can be challenging for orthopedic surgeons, and, as suggested, educating effectively on challenging surgical approaches like the DAA poses even greater difficulties. Therefore, the measurement and evaluation of quantitative indicators of surgical techniques will be an important task for the next generation of surgeons and the safety of patients.

However, the evaluation of competency for these surgeries is not being performed properly in the hospital field. As a method of analyzing and evaluating the learning curve of a surgical technique, it is difficult to represent the learning curve prediction of surgeons because only the operation time and the frequency of complications are used.[5,6] To solve these problems, the entire process of surgery must be classified based on the video of the surgery, and each classified technique must be evaluated.

Recently, technologies such as computer vision, machine learning, and deep learning are rapidly developing.[7-9] In particular, for video channels such as YouTube, video indexing technology using artificial intelligence (AI) is developing and becoming more accurate.[10,11]

Therefore, the purpose of this study was to classify THA videos using deep learning algorithms and evaluate the accuracy of the labelling of the classified videos. Furthermore, the study aimed to train and assess AI that can automatically classify and recognize surgical instruments and surgical steps used in THA procedures when applied to real-time video footage.

METHODS

This study adhered to the principles of the Declaration of Helsinki and was approved by the Institutional Review Board of Gyeongsang National University Hospital (No. GNUH 2019-05-018-01). All individuals included in this study have given their informed consent for the publication of the results.

Data and Reference Standard

Our training dataset consisted of videos of 10 THA surgeries performed by faculty surgeons and recorded between July 2019 and June 2020. The surgical approach for THA performed in this study was a posterolateral approach. To verify the results, we used 2 different datasets: an external dataset and an internal dataset. The first dataset was made up of education videos from the American Academy of Orthopaedic Surgeons (AAOS) website and the second dataset was our own learning data. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline.

Surgical Video Recording Protocol for Training and Test Set

Focusing on the distance from which the surgical field and surgical equipment can be shot, the camera (EOS 100d 18–55 mm, Canon) was fixed to the operator's head using Jimmy Jib (YB-K275, Horusbennu) and then recorded the video. When recording video, the camera angle was maintained at 30° as shown in Fig. 1, and filming was conducted.

Annotation for Surgical Procedure

Using prespecified definitions for the 7 phases, 1 surgeon (JIY) annotated the start and end of the phases in each video. Manual annotations were performed in 7 phases in THA: skin incision, broaching, exposure of acetabulum, acetabular reaming, acetabular cup positioning, real stem insertion, and skin closure (Fig. 2). The total number of captured pictures was 210, with 30 pictures used for each phase.

Fig. 1. Surgical video recording system.

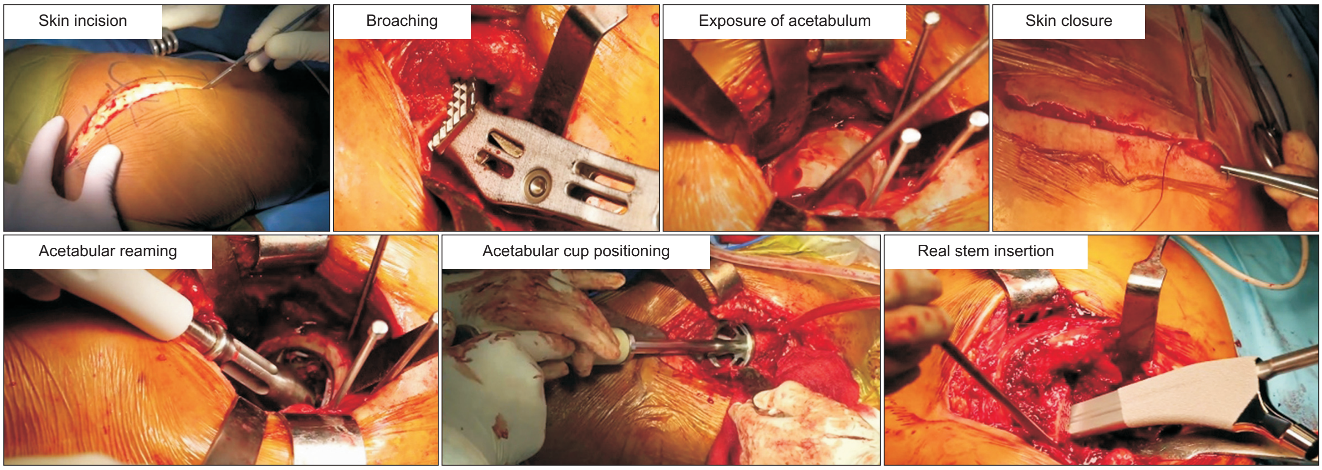

Fig. 2. Seven phases in total hip arthroplasty.

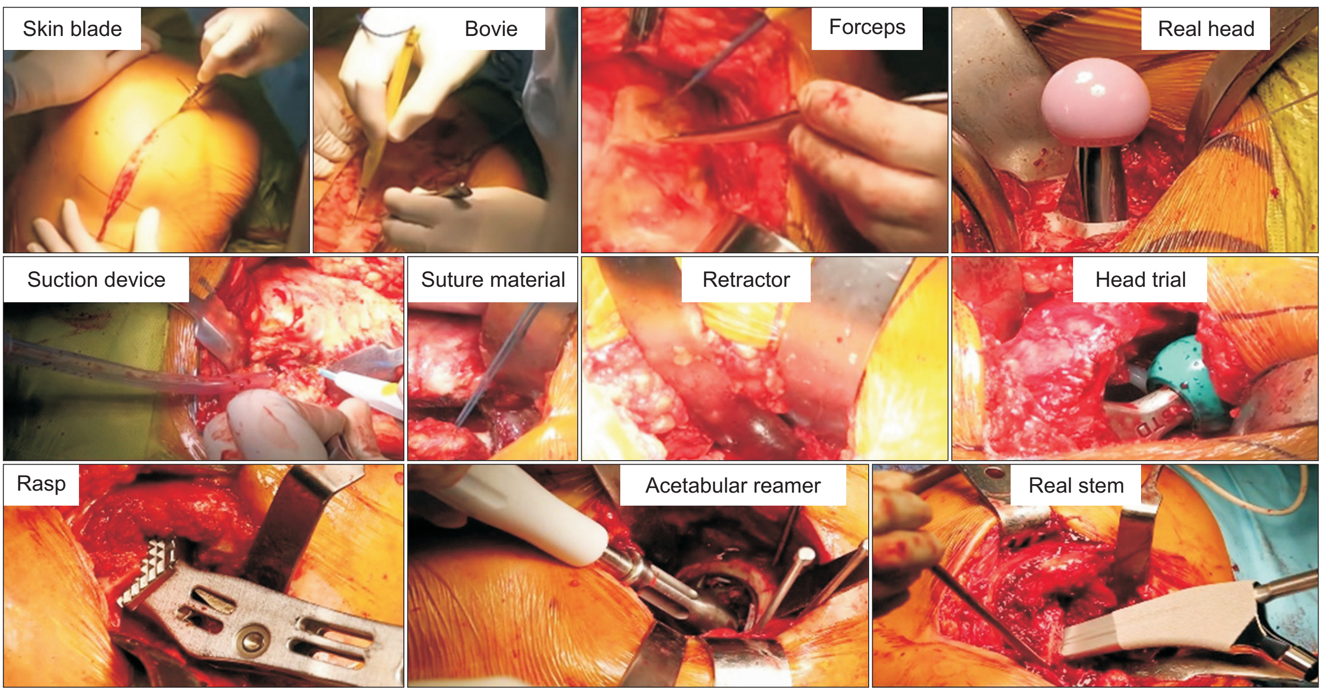

Fig. 3. Eleven Surgical instruments for total hip arthroplasty.

Annotation for Surgical Instruments

Within each phase, another trained annotator (JIY) marked the beginning and end of the use of the following instruments: skin blade, forceps, Bovie, suction device, suture material, retractor, rasp, real stem, acetabular reamer, head trial, real head (Fig. 3). The total number of captured pictures is 330, with 30 images used for each equipment. For the annotation of operational phases, we made bounding boxes on the representative scenes (e.g., skin incision phase can be represented by a long-stretched skin cut). For the annotation of operational tools, we also made bounding boxes on the corresponding tools. Annotation of bounding boxes was implemented by software Labelling.

Model Training and Validation for Object Detection

Using the object bounding box annotations, we implemented the YOLO3 model for the training.[12] For the implementation of the training, Keras platform was used (https://keras.io/). The predicted bound box labels and the corresponding time point of selected scenes were extracted, and their combinations were converted to infer the operational phases. The inferred operational phases and the time points were recorded on the original input video based on in-house Python code.

Firstly, we downloaded 540 operation capture images of THA procedures from internal surgical videos to build a stem detection model using YOLOv3. It consisted of 330 images of surgical instruments and 210 images of surgical procedures. Manually labelled images were successfully used to train the video indexing model. To generate a cross-section of labels for each phase instance, we sampled frames with a unique combination of instruments used during the phase (https://github.com/tzutalin/labelImg). The absence of task-specific images has been a significant impediment in the development and training of recognition models. To increase the training data, we applied image augmentations such as histogram equalization, flipping, and rotation to the original captured operation images, resulting in augmented images. These augmented images were subjected to our pre-trained video indexing model to crop operation procedure fields.

Using a simple Convolution Neural Network (CNN) architecture consisting of 6 layers, we developed a feature extractor based on these THA procedure images. This extractor was able to effectively cluster query images into known THA procedures. The input to the CNN was a set of 224 × 224 grayscale images, which were passed through 2 convolution layers and a max pooling layer to generate a feature map. This map was then fed into 2 fully connected layers, which generated class outputs. Finally, validation was performed on 3 external and internal videos.

Seven Phases Model Construction for Video Indexing

A total of 7 phases represented by 11 instruments for THA were predicted from the operation videos.

RESULTS

Tables 1 and 2 are the results of verification using 3 videos of our dataset. Table 1 shows the mean average precision (mAP) results for 7 phases in THA. The relatively clear classification of surgical techniques such as skin incision and closure, broaching, acetabular reaming, and real stem insertion showed relatively high accuracy with mAP of 0.75 or higher. However, in cases where definitive classification was difficult, such as the definition of acetabular exposure and cup position, the mAP was also 0.33, which is less accurate than other classifications. Table 2 shows the mAP results for the 11 instruments for THA. Except for the suction device, suture material, and retractor, most of the equipment showed good accuracy of mAP 0.7 or higher.

The results of the 3 education videos of the AAOS website are shown in Tables 3 and 4. Table 3 shows the mAP results for 7 phases in THA using the external data-

Table 1. Accuracy for Labelling of the Classified 7 Phases in Total Hip Arthroplasty (Training-Validation Set: Same)

	Skin incision	Broaching	Exposure of acetabulum	Acetabular reaming	Acetabular cup positioning	Real stem insertion	Skin closure
Accuracy (mAP)	0.84	0.97	0.33	0.97	0.12	0.75	0.84

mAP: mean average precision.

Table 2. Accuracy for Labelling of the Classified 11 Instruments for Total Hip Arthroplasty (Training-Validation Set: Same)

	Skin blade	Forceps	Bovie	Suction device	Suture material	Retractor	Rasp	Real stem	Acetabular reamer	Head trial	Real head
Accuracy (mAP)	0.97	0.87	0.72	0.33	0.21	0.56	0.80	0.82	0.87	0.70	0.72

mAP: mean average precision.

Table 3. Accuracy for Labelling of the Classified 7 Phases in Total Hip Arthroplasty (Training-Validation Set: Different)

	Skin incision	Broaching	Exposure of acetabulum	Acetabular reaming	Acetabular cup positioning	Real stem insertion	Skin closure
Accuracy (mAP)	0.94	0.71	0.23	0.68	0.10	0.72	0.94

mAP: mean average precision.

Table 4. Accuracy for Labelling of the Classified 11 Instruments for Total Hip Arthroplasty (Training-Validation Set: Different)

	Skin blade	Forceps	Bovie	Suction device	Suture material	Retractor	Rasp	Real stem	Acetabular reamer	Head trial	Real head
Accuracy (mAP)	0.60	0.34	0.40	0.56	0.24	0.21	0.56	0.47	0.50	0.48	0.47

mAP: mean average precision.

set. The relatively clear classification of surgical techniques such as skin incision and closure, broaching, and real stem insertion showed good accuracy with mAP of 0.71 or higher. However, the identification of the acetabular procedure showed low accuracy as in the same test-validation set. Table 4 shows the mAP results for the 11 instruments for THA using the external dataset. In the case of different test-validation sets, the recognition accuracy of surgical instruments was mostly lower than 0.6 (mAP).

DISCUSSION

This study suggests that using deep learning techniques, THA images can be accurately identified with a small amount of data. The relatively clear classification of surgical techniques, in particular, demonstrated relatively high accuracy with both the same and different test-validation sets. Surgical image analysis has several potential areas of study. The use of AI cameras with deep learning algorithms in surgical video recording and analysis has the potential to be enhanced with automatic zoom and focus technology at each stage. This advancement could improve the accuracy and efficiency of the analysis process and improve the viewing experience for medical professionals and students.[13-15] AI cameras equipped with deep learning algorithms have various clinical applications in surgical images. For example, they can be used for computer-assisted surgery by providing real-time analysis and guidance to surgeons.[16-18] Additionally, deep learning-powered analysis of surgical videos can be utilized in the training and education of medical students and professionals.[13,14,19] Furthermore, these algorithms can aid in the diagnosis and treatment planning of diseases by analyzing surgical images.[20] It is possible that there are additional, unexplored clinical applications for AI cameras in surgical images. As technology advances, it is likely that we will see more ways in which it can be used to enhance patient care and outcomes.[15,18]

The second potential area of research involves the automatic organization of surgical records. As an example, deep learning-based automatic surgical record organization utilizes machine learning algorithms to analyze and categorize surgical records and documents.[19,21,22] One potential benefit of this approach is an improvement in the efficiency of medical record keeping and information management through the automatic classification and organization of surgical records.[19,22] This can allow medical professionals to more easily access and review relevant information, potentially saving time and resources.[19] Additionally, automatic surgical record organization using deep learning may be able to identify patterns and trends in surgical data, which could be useful for research and analysis.[23] For instance, machine learning algorithms could analyze a large number of surgical records to identify common complications or effective treatment strategies.

Another area of study could involve the automatic calculation of blood loss within the surgical field.[24,25] Bleeding during surgery is a key factor in patient prognosis, and using image analysis to objectively track blood loss over time could be valuable.[24,25] Additionally, it is worth exploring the potential applicability of this AI model to different surgical approaches, which could expand its utility and impact in various surgical settings.

In addition, there are several other potential research areas. One of these is the measurement of time for each surgical procedure. Operation time is a crucial health indicator, particularly for elderly patients whose anesthesia time is influenced by the length of the surgery and can affect their recovery.[13,26] Automating intraoperative time tracking can help reduce surgical time, standardize operating room appointment times, and reduce patient costs.[26] In addition, analyzing the time spent in each stage of similar surgeries can provide insights into the relationships between those stages.

Lastly, research on learning curves could be conducted using surgical image analysis.[13] It is believed that it takes at least dozens of surgeries to become proficient in a particular skill, and tracking where the time required for a skill decreases through image analysis can be useful in understanding the learning curve.

This study has some limitations. The first is the small sample size. In this study, 10 videos were used, and in the process, the number of samples was expanded by applying various conversion techniques. Since learning was conducted in a state where the number of initial images was small, a process to further improve performance was additionally required. If the number of initial images is sufficient, high performance can be obtained with only existing images, and richer verification and test data sets can be built when various conversion techniques are additionally applied.

The second limitation can be grouped into environmental factors. During the process of deep learning, the recognition rate may decrease when the proximity or external screen is dark or when similar procedures such as skin incision and closure are recognized incorrectly. To improve the recognition rate in these situations, the data images must be rearranged according to the frame sequence. Additionally, the recognition rate can be improved by recognizing important surgical equipment and surrounding procedures together.[16] Therefore, it is crucial to select important tools and scenes for each procedure and include them in the learning process to increase accuracy.[16]

In addition to the mentioned limitations, it is essential to consider the complexity of surgical procedures and the diversity of surgical instruments when discussing the performance of AI models in real-world surgical applications. As observed in our study, different surgical procedures exhibit varying levels of complexity, and this intricacy can pose challenges for the accurate classification of surgical steps and instruments by AI models. The performance of these models may depend on their ability to comprehend and distinguish the nuances between the steps and instruments associated with each procedure. Moreover, the diversity in surgical instruments, characterized by differences in shapes and sizes, as well as their varied usage across different procedures, is a critical factor influencing AI model performance. During our research, we noted that the model's exposure to a broad spectrum of instruments during training significantly impacted its ability to recognize and classify these instruments accurately during testing scenarios. The limitations associated with instrument recognition may arise if the model has not encountered certain instruments during its training phase.

Furthermore, factors such as shooting distance, angle, lighting, training environment, and tool shape can significantly impact deep learning performance.[24,25] As a result, it is crucial to gather high-quality data through the use of an auto-focusing function and technique that can consistently maintain the appropriate shooting distance, angle, and brightness.[27] Additionally, the factors of shooting distance, angle, lighting, training environment, and tool shape all influence the effectiveness of deep learning. As such, it is essential to collect high-quality data through the use of an auto-focusing function and method that can consistently maintain the proper shooting distance, angle, and brightness.[27] Endoscopic surgery conducted in a closed field seems to be more readily classified; however, external factors must be fully taken into account when performing surgery in an open field. Further research will be necessary to address this issue in the future.

The feasibility and accuracy of using deep learning techniques for the automated identification of surgical phases in THA videos were demonstrated in our study. Our results showed that the femoral stem selection in THA patients could be identified with a high degree of accuracy, with mAP values of 0.7 or higher, except for a specific step in the 2 validation datasets. Therefore, our proposed method can be a reliable and efficient alternative to the traditional manual method, which may help improve surgical efficiency, reduce the risk of errors, and ultimately lead to better patient outcomes.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2022R1C1C1004134).

ORCID

Yang Jae Kang https://orcid.org/0000-0002-4313-7088
Shin June Kim https://orcid.org/0009-0004-9570-7830
Sung Hyo Seo https://orcid.org/0000-0003-3327-5425
Sangyeob Lee https://orcid.org/0000-0001-8024-1135
Hyeon Su Kim https://orcid.org/0000-0002-6883-5448
Jun-Il Yoo https://orcid.org/0000-0002-3575-4123

REFERENCES

1. Grau L, Summers S, Massel DH, Rosas S, Ong A, Hernandez VH. Operative trends in the treatment of hip fractures and the role of arthroplasty. Geriatr Orthop Surg Rehabil. 2018;9:2151459318760634.
2. Gasbarra E, Perrone FL, Celi M, et al. Total hip arthroplasty revision in elderly patients. Aging Clin Exp Res. 2013;25 Suppl 1:S61-3.
3. Villanueva-Martınez M, Hernandez-Barrera V, Chana-Rodriguez F, et al. Trends in incidence and outcomes of revision total hip arthroplasty in Spain: a population based study. BMC Musculoskelet Disord. 2012;13:37.
4. Kayani B, Konan S, Ayuob A, Ayyad S, Haddad FS. The current role of robotics in total hip arthroplasty. EFORT Open Rev. 2019;4(11):618-25.
5. Papachristofi O, Jenkins D, Sharples LD. Assessment of learning curves in complex surgical interventions: a consecutive case-series study. Trials. 2016;17(1):266.
6. Maruthappu M, Duclos A, Lipsitz SR, Orgill D, Carty MJ. Surgical learning curves and operative efficiency: a cross-specialty observational study. BMJ Open. 2015;5(3):e006679.
7. Chellappa R, Theodoridis S, Van Schaik A. Advances in machine learning and deep neural networks. Proc IEEE. 2021;109(5):607-11.
8. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. IEEE Access. 2019;7:53040-65.
9. Salas J, de Barros Vidal F, Martínez-Trinidad F. Deep learning: current state. IEEE Lat Am Trans. 2019;17(12):1925-45.
10. Hu W, Xie N, Li L, Zeng X, Maybank S. A survey on visual content-based video indexing and retrieval. IEEE Trans Syst Man Cybern C Appl Rev. 2011;41(6):797-819.
11. Mahrishi M, Morwal S, Muzaffar AW, Bhatia S, Dadheech P, Rahmani MK. Video index point detection and extraction framework using custom YoloV4 Darknet object detection model. IEEE Access. 2021;9:143378-91.
12. Redmon J, Farhadi A. Yolov3: an incremental improvement. arXiv [Preprint]. 2018 [cited 2023 Nov 12]. Available from: https://doi.org/10.48550/arXiv.1804.02767
13. Takeuchi M, Collins T, Ndagijimana A, et al. Automatic surgical phase recognition in laparoscopic inguinal hernia repair with artificial intelligence. Hernia. 2022;26(6):1669-78.
14. Huaulme A, Despinoy F, Perez SA, Harada K, Mitsuishi M, Jannin P. Automatic annotation of surgical activities using virtual reality environments. Int J Comput Assist Radiol Surg. 2019;14(10):1663-71.
15. Kutana S, Bitner DP, Addison P, Chung PJ, Talamini MA, Filicori F. Objective assessment of robotic surgical skills: review of literature and future directions. Surg Endosc. 2022;36(6):3698-707.
16. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. Surg Endosc. 2020;34(11):4924-31.
17. Sanchez-Matilla R, Robu M, Grammatikopoulou M, Luengo I, Stoyanov D. Data-centric multi-task surgical phase estimation with sparse scene segmentation. Int J Comput Assist Radiol Surg. 2022;17(5):953-60.
18. Ershad M, Rege R, Majewicz Fey A. Automatic and near real-time stylistic behavior assessment in robotic surgery. Int J Comput Assist Radiol Surg. 2019;14(4):635-43.
19. Bombieri M, Rospocher M, Dall'Alba D, Fiorini P. Automatic detection of procedural knowledge in robotic-assisted surgical texts. Int J Comput Assist Radiol Surg. 2021;16(8):1287-95.
20. Dergachyova O, Bouget D, Huaulme A, Morandi X, Jannin P. Automatic data-driven real-time segmentation and recognition of surgical workflow. Int J Comput Assist Radiol Surg. 2016;11(6):1081-9.
21. Garrow CR, Kowalewski KF, Li L, et al. Machine learning for surgical phase recognition: a systematic review. Ann Surg. 2021;273(4):684-93.
22. Huaulme A, Jannin P, Reche F, Faucheron JL, Moreau-Gaudry A, Voros S. Offline identification of surgical deviations in laparoscopic rectopexy. Artif Intell Med. 2020;104:101837.
23. Guedon AC, Meij SE, Osman KN, et al. Deep learning for surgical phase recognition using endoscopic videos. Surg Endosc. 2021;35(11):6150-7.
24. Garcia-Martinez A, Vicente-Samper JM, Sabater-Navarro JM. Automatic detection of surgical haemorrhage using computer vision. Artif Intell Med. 2017;78:55-60.
25. Jo K, Choi B, Choi S, Moon Y, Choi J. Automatic detection of hemorrhage and surgical instrument in laparoscopic surgery image. Annu Int Conf IEEE Eng Med Biol Soc. 2016;2016:1260-3.
26. Zafar SU, Khan FA, Khan M. Standardization of anaesthesia ready time and reasons of delay in induction of anaesthesia. J Pak Med Assoc. 2006;56(3):112-5.
27. Cho SM, Kim YG, Jeong J, Kim I, Lee HJ, Kim N. Automatic tip detection of surgical instruments in biportal endoscopic spine surgery. Comput Biol Med. 2021;133:104384.