SurgeonAssist-Net: Towards Context-Aware Head-Mounted Display-Based Augmented Reality for Surgical Guidance

Mitchell Doughty[1,2(✉)], Karan Singh[3], and Nilesh R. Ghugre[1,2,4]

[1] Department of Medical Biophysics, University of Toronto, Toronto, Canada
mitchell.doughty@mail.utoronto.ca
[2] Schulich Heart Program, Sunnybrook Health Sciences Centre, Toronto, Canada
[3] Department of Computer Science, University of Toronto, Toronto, Canada
[4] Physical Sciences Platform, Sunnybrook Research Institute, Toronto, Canada

Abstract. We present SurgeonAssist-Net: a lightweight framework making action-and-workflow-driven virtual assistance, for a set of predefined surgical tasks, accessible to commercially available optical see-through head-mounted displays (OST-HMDs). On a widely used benchmark dataset for laparoscopic surgical workflow, our implementation competes with state-of-the-art approaches in prediction accuracy for automated task recognition, and yet requires 7.4× fewer parameters, 10.2× fewer floating point operations per second (FLOPS), is 7.0× faster for inference on a CPU, and is capable of near real-time performance on the Microsoft HoloLens 2 OST-HMD. To achieve this, we make use of an efficient convolutional neural network (CNN) backbone to extract discriminative features from image data, and a low-parameter recurrent neural network (RNN) architecture to learn long-term temporal dependencies. To demonstrate the feasibility of our approach for inference on the HoloLens 2 we created a sample dataset that included video of several surgical tasks recorded from a user-centric point-of-view. After training, we deployed our model and cataloged its performance in an online simulated surgical scenario for the prediction of the current surgical task. The utility of our approach is explored in the discussion of several relevant clinical use-cases. Our code is publicly available at https://github.com/doughtmw/surgeon-assist-net.

Keywords: Augmented reality · Machine learning · Surgical task prediction · Head-mounted display · Microsoft HoloLens · Neural networks

1 Introduction

There has been significant interest in adoption of augmented reality (AR) for surgical guidance in the medical field, due to its ability to enhance task performance when effectively implemented [1]. The use of a see-through head-mounted

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12904, pp. 667–677, 2021.
https://doi.org/10.1007/978-3-030-87202-1_64

display (HMD) as the visualization medium, as opposed to a monitor, has been demonstrated to provide a further benefit to efficiency by eliminating the visual disconnect between the monitor and the surgical scene [2].

Though recent work has indicated the applicability of AR in laparoscopic and endoscopic procedures [3,4], neurosurgery [5], orthopedic surgery [6], and general surgery [7], there remains the concern of overloading the user with too much additional information, distracting them from the task at hand and resulting in reduced performance over routine standardized techniques [8].

Motivation. The bulk of research work into AR systems for medical image-guidance has centered around technical developments, including calibration [9], alignment [1], and visualization [10,11] and focused on achieving quantitative metrics like speed and accuracy [12]. Due to a lack of optimized virtual workflow and information representation, these advances do not guarantee the effective translation of guidance systems to clinical practice; this is evidenced by the absence of widely used commercial see-through HMD navigation systems [13].

To bridge the gap towards an effective OST-HMD based AR guidance system, our aim was to address these pitfalls by creating a framework capable of understanding the current action of a user and supporting them with only the critical information that is relevant to the task at hand, thus reducing the information overload and the need for manual control of displayed virtual content.

Related Work. Context-aware surgery involves the interpretation of the large amount of information created during a surgical procedure, with focus on recognizing/predicting key tasks [14], monitoring incidents [15], and highlighting adverse events. Surgical task prediction in an off-line context has been recently investigated in neurosurgery [16], laparoscopy [17], and cataract surgery [18] and has proposed the use of a secondary monitor to display virtual assistance. These applications have relied on various types of input features to predict the current surgical phase, such as radio-frequency identification chips attached to surgical instruments [17], instrument signals [19], robot kinematic data [20], external infrared measurement systems [21], or laparoscopic video [14].

Recent applications to context-aware surgery have remained limited to procedures where high-performance computing resources and video data are readily accessible [14,22,23]. If these systems are to be generalized to other surgical procedures, the display of context-aware information on a secondary monitor could introduce a potentially disorienting visual disconnect for the user.

To address these challenges, Katić et al. propose a context-aware system, based on an OST-HMD, for intraoperative AR in dental implant surgery. In a porcine study, the authors demonstrated an improved task completion time and acceptance of their system by dental surgeons [21]. However, the reliance on additional sensors, computing power and specialized markers for task prediction makes their method challenging to incorporate into a typical operating room workflow and incapable of generalization to different surgical scenarios.

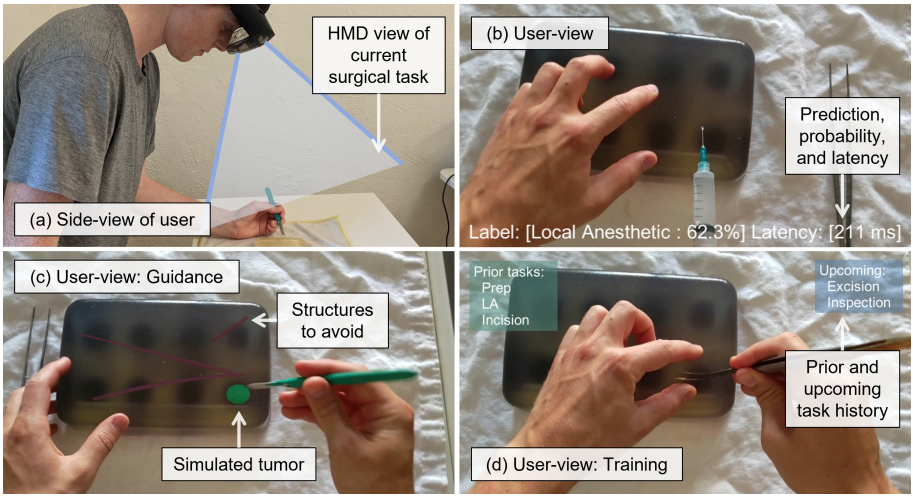

Fig. 1. Example images from online inference with the SurgeonAssist-Net app on a HoloLens 2. (a) Side-view of the user and phantom. User-view with (b) information on the current surgical phase prediction, (c) minimal virtual models for task-specific guidance, and (d) information on prior and upcoming surgical tasks for user-training.

In contrast to these systems, we propose SurgeonAssist-Net, a novel framework to predict the current surgical task that a user is performing and, using the context-aware predictions, ensure that the virtual augmentation meets the current information needs of the user. Our approach does not rely on the use of external sensors, custom/specialized hardware, or additional computing power. We are the first to demonstrate the implementation of a context-aware platform on a commercially available OST-HMD with near real-time performance, eliminating the visual disconnect between a monitor and the patient (Fig. 1).

2 Methods

2.1 SurgeonAssist-Net: Surgical Task Prediction

To provide context-awareness to the wearer of an OST-HMD from user-centric input video, we have created the SurgeonAssist-Net framework, composed of an EfficientNet-Lite-B0 [24,25] backbone for feature extraction and a gated recurrent unit (GRU) RNN framework [26] for storing long term dependency information (Fig. 2). These networks are jointly trained in an end-to-end manner, generating features that encode both spatial and temporal information.

Spatial Feature Extraction. Deep learning and the introduction of deep CNNs has led to vast improvements in interpreting high-dimensional data over traditional approaches, finding successful applications in object detection and

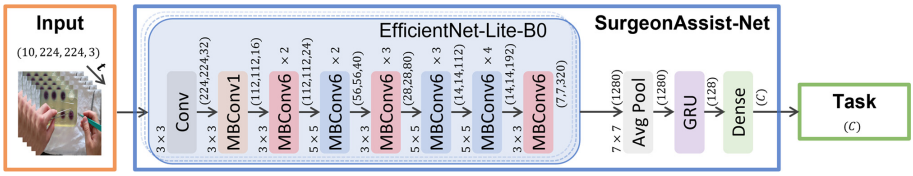

Fig. 2. Overview of the deep learning-based framework for extracting relevant information from video frame data (10 frames) and predicting the current surgical task.

image recognition [27]. With EfficientNet, Tan et al. overcame scaling issues common to increasingly deep CNNs through a compound scaling method that optimally adjusted the width, depth, and resolution of the network by using fixed coefficients, thus achieving a balance between network speed and accuracy [24]. On ImageNet, EfficientNet-B0 outperforms ResNet-50 [28] in top-1 and top-5 accuracy and offers a 4.9× parameter and 10.5× FLOPS reduction [24].

With EfficientNet-Lite-B0 [25], modifications to the EfficientNet-B0 model were implemented to optimize performance for mobile CPU applications; we use pre-trained weights from ImageNet to serve as an initial starting point for training. The final fully connected layer at the end of the EfficientNet-Lite-B0 network was removed and replaced with a global average-pooling layer to output a $7 \times 7 \times 320$ tensor of high-level discriminative features that was reshaped to a vector of length 1280 to serve as an input to the GRU framework (Fig. 2).

Temporal Information Modeling. Recurrent neural networks can handle variable-length sequence inputs by using a hidden state augmented with nonlinear mechanisms whose activation at each time step is reliant on that of the prior frame [29]. Both GRU and long short-term memory (LSTM) components have been used for time-series forecasting tasks like analysis of video data for activity recognition, image captioning, and surgical task prediction [14]. It has been demonstrated that GRUs perform similarly to LSTM units [30], but have fewer total parameters and are more well-suited to real-time inference applications.

For our RNN architecture, we found optimal results using a single GRU cell with 128 hidden units, followed by a decision network with a ReLU activation, dropout layer with probability of 0.2, and a fully connected output layer with C output nodes—corresponding to the C potential surgical tasks. The parameters of the GRU cell and dense layer were initialized using Xavier normal initialization [31]. During inference, we used an online recognition mode accessing only current and prior frames. After performing initial hyperparameter evaluation experiments, we found that a sequence length of 10 video frames provided an optimal trade-off for system performance and speed (Fig. 2).

2.2 Integrating SurgeonAssist-Net for Online Inference

We used the Microsoft HoloLens 2 OST-HMD for the recording of user-centric video and visualization of context-aware surgical task predictions. The HoloLens 2 is capable of visualization of three-dimensional (3D) virtual models through stereoscopic vision via two-dimensional (2D) laser beam scanning displays, offering a field of view of 43×29 degrees (horizontal $\times$ vertical) to the wearer.

Input frames of size 896×504 px were requested from the HoloLens 2 photo-video camera at a rate of 30 frames per second (FPS), resized to 256×256 px using nearest-neighbor interpolation [32], and center cropped to a final resolution of 224×224 px; these served as input to the prediction framework. We leveraged the Windows Machine Learning and Open Neural Net Exchange (ONNX) [33] libraries within a C# Universal Windows Platform (UWP) app to perform inference using the SurgeonAssist-Net model. The OpenCV library [32] was included within a C++ UWP runtime component to prepare input frame data for prediction. The network output, the predicted task, was then used to optimize the virtual content shown to the user based on their current information needs.

2.3 Cholec80 Dataset

Cholec80 contains 80 videos (1920×1080 px or 854×480 px at 25 FPS) of cholecystectomy surgeries performed by 13 surgeons, complete with phase annotations of the 7 surgical phases for a procedure (25 FPS) defined by a senior surgeon [14]. The original videos were down-sampled from 25 FPS to 1 FPS to match the temporal resolution used by other groups [23]. We use nearest-neighbor interpolation to scale the input video frames from the original resolution to 256×256 px to improve computational efficiency [32]. For all tests using the Cholec80 dataset, 32 videos were used for a train split, 40 videos for a test split, and the remaining 8 videos for a validation split, as in prior work [14].

Our framework was implemented using the PyTorch [34] deep learning library. We trained our network for 25 epochs with a batch size of 32 on $2\times$ NVIDIA V100 GPUs, each with 32 GB HBM2 memory, and reported the average network performance across three training runs. For optimization, we used stochastic gradient descent (SGD), an initial learning rate of $5e^{-4}$, and a momentum of 0.9. Sequence-wise data augmentation was performed on each training batch of image data, including random cropping of input images to 224×224 px, horizontal and vertical flipping, and random color augmentation.

To evaluate the performance of the SurgeonAssist-Net for task recognition, we employed the widely used metrics of accuracy (AC), precision (PR), and recall (RE) [14] and compared the results to other recent approaches. Furthermore, we included an estimate of the total number of parameters in each model, the model size, the inference time (latency), and the FLOPS for an input image sequence of size $(t \times 224 \times 224 \times 3)$, where t is the input sequence length of that specific approach. A single core of an AMD Ryzen 3600 CPU was used to measure the average latency for each network over 10 runs on a subset of the testing data.

2.4 User-Centric Surgical Tasks Dataset

Due to the lack of available video of surgical tasks from a user-centric point of view, we created our own dataset for training the SurgeonAssist-Net framework and evaluating its online performance on the HoloLens 2 device. The dataset included a total of five surgical tasks performed by three novice users on a gelatin phantom as they worked to remove a simulated subsurface tumor. During the task, the typical suite of surgical tools: scalpel, forceps, scissors, clamp, and syringe, was made available to the users. We recorded the dataset using the photo-video camera on the HoloLens 2 OST-HMD (1280×720 px at 30 FPS). A total of $52,500$ frames were extracted from the videos at a rate of 30 FPS. Details of the dataset including tasks and duration are included in Table 1.

Table 1. Details of the user-centric surgical tasks dataset.

Phase	Preparation	Local anes.	Incision	Excision	Inspection
Duration (sec)	27.2 ± 11.5	39.2 ± 10.0	15.1 ± 9.9	32.3 ± 23.5	12.3 ± 8.9
Annotations	$4,110$	$5,880$	$13,800$	$21,360$	$7,350$

As with the Cholec80 benchmark dataset, we resized input video frames of the user-centric surgical tasks dataset to a resolution of 256×256 px [32] and used an input sequence length of 10 frames. We segmented the dataset such that 3 videos were used for a train split, 1 video for a test split, and the remaining 1 video for a validation split.

3 Results and Discussion

3.1 Cholec80: Surgical Task Prediction on a Benchmark Dataset

Table 2 compares the performance and latency of each approach on the Cholec80 dataset. Twinanda et al. have presented surgical task prediction results using learned visual features and temporal dependencies [14] based on (1) the single-task PhaseNet with features extracted from a modified AlexNet backbone [35]; and (2) the multi-task EndoNet framework which makes use of a modified AlexNet backbone for feature extraction and tool classification; both approaches use a single image frame as the input sequence. The single-task SV-RCNet [22] and multi-task MTRCNet-CL [23] networks share a similar ResNet-50 architecture for feature extraction and an LSTM cell with 512 hidden nodes for phase prediction. Additionally, both the SV-RCNet and MTRCNet-CL approaches use an input sequence length of 10 frames for prediction. As we were only interested in single-task performance, we did not report the results of multi-task approaches like EndoNet and MTRCNet-CL in our evaluation.

SurgeonAssist-Net outperformed the PhaseNet [14] framework across AC, PR, and RE metrics, required $46\times$ fewer parameters for operation, and used

45× less memory for model deployment. When compared to SV-RCNet [22], SurgeonAssist-Net scored better in AC and PR metrics and required 7.4× fewer model parameters, achieved 10.2× faster FLOPS, and used 3× less time for inference. Due to its performance efficiency, low parameter count, and compact model size, SurgeonAssist-Net can be effectively operated in computationally restricted environments for real-time inference. Figure 3 provides a qualitative representation of the performance of SurgeonAssist-Net, without any form of post-processing, across a subset of the Cholec80 dataset (Video41).

Table 2. Results versus state-of-the-art using the Cholec80 surgical tasks dataset.

Method	Frames (t)	AC (%)	PR (%)	RE (%)	Size (MB)	Params (M)	FLOPS (B)	Latency (ms)
Ours	10	**85.8**	**81.5**	81.4	**15.9**	**3.91**	4.04	532.4
SV-RCNet [22]	10	85.3	80.7	**83.5**	115.3	28.76	41.25	1593.8
PhaseNet [14]	1	73.0	67.0	63.4	718.8	179.71	**0.83**	**99.3**

Fig. 3. SurgeonAssist-Net phase prediction (Pred) performance visualized relative to the ground truth (GT) phase labels on Video41 of the Cholec80 dataset (51 m 43 s in duration). The legend indicates the color coding of each individual phase.

3.2 User-Centric Surgical Tasks Dataset: Task Prediction and Online Performance on the HoloLens 2

To make the SurgeonAssist-Net model available to the HoloLens 2 through the UWP interface, we converted our trained model from its PyTorch implementation to ONNX format [33]. Table 3 includes the relative performance of the SurgeonAssist-Net framework in PyTorch and ONNX formats compared with ONNX converted implementations of PhaseNet [14] and SV-RCNet [22] when evaluated on the user-centric surgical tasks dataset.

A small decrease in AC, PR, RE and model size was recorded following conversion of the SurgeonAssist-Net model to ONNX format. However, we also measured a 5.2× decrease in CPU inference time by the ONNX model when compared to its PyTorch implementation; this speedup was due to the compilation of an efficient graph model during ONNX model export. Similar relative performance across AC, PR, and RE by the SurgeonAssist-Net model as compared to the ONNX converted PhaseNet [14] and SV-RCNet [22] was observed.

HoloLens 2 Performance and Feasibility. To evaluate the real-world performance of the SurgeonAssist-Net model on the HoloLens 2 headset, we created a sample application that displayed the prediction, with its associated probability and latency, for the current surgical task being performed. In Fig. 1 we include a sample image from an experiment where a user was presented with the same gelatin phantom and surgical tools as in the user-centric surgical tasks dataset and tasked with removing a subsurface tumor. Across the test, there was good agreement between the predicted and user-performed tasks.

The latency of the SurgeonAssist-Net ONNX model on the HoloLens 2 CPU, averaged across 30 s of online predictions, was measured to be 219.2 ms, or roughly 5 FPS, with a single image input sequence. To measure the feasibility of online inference with other networks on the HoloLens 2, we repeated the above experiment with ONNX converted PhaseNet [14] and SV-RCNet [22] models; however, we were unable to successfully load or operate either model on the HoloLens 2 CPU due to the large model size and/or high FLOPS requirements.

Table 3. Results versus state-of-the-art using the user-centric surgical tasks dataset.

Method	Frames (t)	AC (%)	PR (%)	RE (%)	Size (MB)	Latency (ms)
Ours PyTorch	10	**85.5**	**88.4**	76.5	15.9	538.6
Ours ONNX	10	85.1	87.5	75.3	**15.2**	103.1
SV-RCNet ONNX [22]	10	84.9	87.3	**77.5**	112.2	721.2
PhaseNet ONNX [14]	1	70.6	69.7	60.1	702.0	**64.7**

3.3 Clinical Significance

Aside from accuracy, a primary limitation of AR-guided approaches is the reliance upon a user to manually control the appearance and presentation of virtually augmented entities, thereby adapting the visualization to their current surgical context. This is tedious and detracts from their focus on the task. Our work on surgical task prediction is thus critical and foundational in ensuring that the automated augmentation of virtual models meets the current information needs. In this work, the predicted surgical task serves as a prerequisite to displaying the optimal virtual information to the user. We will now briefly discuss three clinical scenarios involving surgical guidance, user-training, and performance evaluation, where the SurgeonAssist-Net could be readily incorporated.

Guidance. The predicted task context can control the choice and presentation of the augmented virtual models. For example, in general surgery, as a surgeon picks up a scalpel and the incision phase of a procedure is detected, the HMD would display a relevant virtual model indicating the target site for surgical entry (Fig. 1). Throughout the procedure, our surgical phase detection would enable different virtual models and information relevant to the surgical phase to be optimally selected and presented, without user intervention.

Training. Task prediction can be used to guide a student, wearing the HMD, in practicing a general surgery task on a cadaver. Their active learning can be reinforced by presenting them with a task history of phases performed, or of upcoming surgical actions, given their present surgical step (Fig. 1). Additional relevant information, in the form of visual cues, text, or audio, could be presented in tandem with the detected task to enhance the training experience.

Evaluation. Task analytics can provide surgeons with quantitative data on a surgical procedure. For example, a surgeon performing a less frequent procedure could wear the HMD while re-training and be provided with chronology and analytics of the time spent in each surgical phase (Fig. 3). This information, when compared to peers, could serve to suggest focus areas for improvement.

4 Conclusions and Future Work

The focus of this work was to create a lightweight framework capable of understanding user-centric activities and providing virtual real-time workflow assistance for a pre-defined set of surgical tasks. By training the SurgeonAssist-Net framework on the user-centric surgical tasks dataset, we were able to use it effectively as an event-detection heuristic to associate known events in an online scenario by using the on-board computing resources of an OST-HMD (in this case, the HoloLens 2). For future investigation, we envision that an in-depth user-study to evaluate a manually-controlled AR experience versus the context-aware approach could further demonstrate the benefits of action-driven virtual assistance. We also expect that a larger user-centric training dataset would result in better consistency in predictions from the SurgeonAssist-Net framework. Nonetheless, we have demonstrated the potential capabilities of an online context-aware surgical guidance platform and brought attention to its capacity to overcome issues which had previously plagued AR-based image-guidance systems.

Acknowledgements. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery program (RGPIN-2019-06367).

References

1. Peters, T.M.: Image-guidance for surgical procedures. Phys. Med. Biol. **51**(14), R505 (2006)
2. Liu, D., Jenkins, S.A., Sanderson, P.M., Fabian, P., Russell, W.J.: Monitoring with head-mounted displays in general anesthesia: a clinical evaluation in the operating room. Anesth. Analg. **110**(4), 1032–1038 (2010)
3. Bernhardt, S., Nicolau, S.A., Soler, L., Doignon, C.: The status of augmented reality in laparoscopic surgery as of 2016. Med. Image Anal. **37**, 66–90 (2017)
4. Zorzal, E.R., et al.: Laparoscopy with augmented reality adaptations. J. Biomed. Inform. **107**, 103463 (2020)

5. Meola, A., Cutolo, F., Carbone, M., Cagnazzo, F., Ferrari, M., Ferrari, V.: Augmented reality in neurosurgery: a systematic review. Neurosurg. Rev. **40**(4), 537–548 (2016). https://doi.org/10.1007/s10143-016-0732-9
6. Jud, L., et al.: Applicability of augmented reality in orthopedic surgery-a systematic review. BMC Musculoskelet. Disord. **21**(1), 1–13 (2020)
7. Rahman, R., Wood, M.E., Qian, L., Price, C.L., Johnson, A.A., Osgood, G.M.: Head-mounted display use in surgery: a systematic review. Surg. Innov. **27**(1), 88–100 (2020)
8. Dixon, B.J., Daly, M.J., Chan, H., Vescan, A.D., Witterick, I.J., Irish, J.C.: Surgeons blinded by enhanced navigation: the effect of augmented reality on attention. Surg. Endosc. **27**(2), 454–461 (2013)
9. Grubert, J., Itoh, Y., Moser, K., Swan, J.E.: A survey of calibration methods for optical see-through head-mounted displays. IEEE Trans. Visual Comput. Graphics **24**(9), 2649–2662 (2017)
10. Kersten-Oertel, M., Jannin, P., Collins, D.L.: The state of the art of visualization in mixed reality image guided surgery. Comput. Med. Imaging Graph. **37**(2), 98–112 (2013)
11. Hong, J., et al.: Three-dimensional display technologies of recent interest: principles, status, and issues [invited]. Appl. Opt. **50**(34), H87–H115 (2011)
12. Cleary, K., Peters, T.M.: Image-guided interventions: technology review and clinical applications. Annu. Rev. Biomed. Eng. **12**, 119–142 (2010)
13. Eckert, M., Volmerg, J.S., Friedrich, C.M.: Augmented reality in medicine: systematic and bibliographic review. JMIR mHealth uHealth **7**(4), e10967 (2019)
14. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**(1), 86–97 (2016)
15. Suzuki, T., Sakurai, Y., Yoshimitsu, K., Nambu, K., Muragaki, Y., Iseki, H.: Intraoperative multichannel audio-visual information recording and automatic surgical phase and incident detection. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 1190–1193. IEEE (2010)
16. Forestier, G., et al.: Multi-site study of surgical practice in neurosurgery based on surgical process models. J. Biomed. Inform. **46**(5), 822–829 (2013)
17. Navab, N., Traub, J., Sielhorst, T., Feuerstein, M., Bichlmeier, C.: Action-and workflow-driven augmented reality for computer-aided medical procedures. IEEE Comput. Graphics Appl. **27**(5), 10–14 (2007)
18. Quellec, G., Lamard, M., Cochener, B., Cazuguel, G.: Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials. IEEE Trans. Med. Imaging **34**(4), 877–887 (2014)
19. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Med. Image Anal. **16**(3), 632–641 (2012)
20. Lea, C., Vidal, R., Hager, G.D.: Learning convolutional action primitives for fine-grained action recognition. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 1642–1649. IEEE (2016)
21. Katić, D., et al.: A system for context-aware intraoperative augmented reality in dental implant surgery. Int. J. Comput. Assist. Radiol. Surg. **10**(1), 101–108 (2014). https://doi.org/10.1007/s11548-014-1005-0
22. Jin, Y., et al.: SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging **37**(5), 1114–1126 (2017)
23. Jin, Y., et al.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med. Image Anal. **59**, 101572 (2020)

24. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
25. Liu, R.: Higher accuracy on vision models with efficientnet-lite. TensorFlow Blog (2020). https://blog.tensorflow.org/2020/03/higher-accuracy-on-visionmodels-with-efficientnet-lite.html. Accessed 30 Apr 2020
26. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
27. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
30. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
31. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
32. Bradski, G.: The opencv library. Dr. Dobb's J. Softw. Tools **25**, 120–125 (2000)
33. Bai, J., Lu, F., Zhang, K., et al.: ONNX: open neural network exchange (2019). https://github.com/onnx/onnx
34. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)
35. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)