The AI-Medic: A Multimodal Artificial Intelligent Mentor for Trauma Surgery

Edgar Rojas-Muñoz
emuoz@purdue.edu
School of Industrial Engineering
West Lafayette, Indiana

Kyle Couperus
kyle.s.couperus.mil@mail.mi
Madigan Army Medical Center
Joint Base Lewis-McChord
Washington

Juan P. Wachs
jpwachs@purdue.edu
School of Industrial Engineering
West Lafayette, Indiana

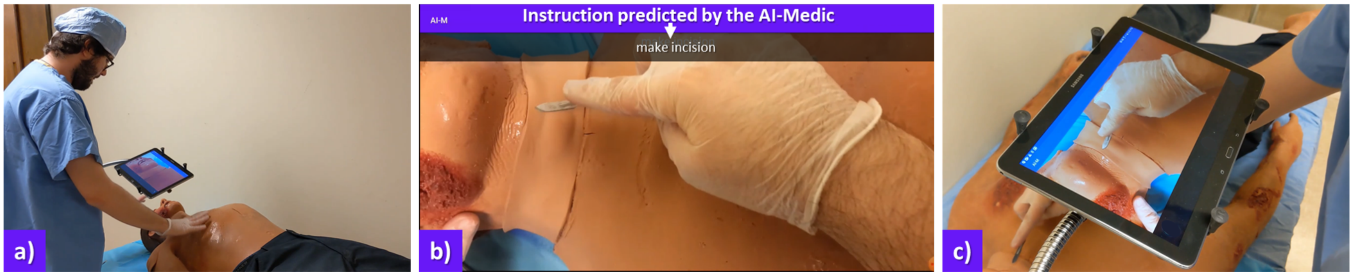


Figure 1: The AI-Medic provides guidance to a surgeon performing a medical procedure (a). The system captures the view of the operating field and predicts the instruction to perform (b). The instruction is conveyed to the surgeon using text-to-speech routines and by displaying it in the screen of a tablet device (c).

ABSTRACT

Telementoring generalist surgeons as they treat patients can be essential when in situ expertise is not readily available. However, adverse cyber-attacks, unreliable network conditions, and remote mentors' predisposition can significantly jeopardize the remote intervention. To provide medical practitioners with guidance when mentors are unavailable, we present the AI-Medic, the initial steps towards the development of a multimodal intelligent artificial system for autonomous medical mentoring. The system uses a tablet device to acquire the view of an operating field. This imagery is provided to an encoder-decoder neural network trained to predict medical instructions from the current view of a surgery. The network was training using DAISI, a dataset including images and instructions providing step-by-step demonstrations of surgical procedures. The predicted medical instructions are conveyed to the user via visual and auditory modalities.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**;
• **Applied computing** → **Health care information systems**.

KEYWORDS

datasets; neural networks; telementoring; surgery

ACM Reference Format:
Edgar Rojas-Muñoz, Kyle Couperus, and Juan P. Wachs. 2020. The AI-Medic: A Multimodal Artificial Intelligent Mentor for Trauma Surgery. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3382507.3421167

1 INTRODUCTION

Telementoring techniques have been explored to provide generalist surgeons with remote supervision when no expert specialist is available on-site [5]. Two aspects are fundamental requirements for these techniques: the availability of the expert that provides medical guidance, and having a reliable communication medium. Nonetheless, such requirements cannot always be satisfied. A possible approach to convey guidance during such situations is to incorporate Artificial Intelligence (AI) into telementoring systems [1]. However, AI-based frameworks for autonomous medical mentoring have been limited due to the lack of robust predicting models, size, and quality of datasets that are required to train such models.

The novel DAISI dataset addresses this gap by compiling images and text descriptions that provide step-by-step demonstrations of how to complete medical procedures [4]. Our work levarages DAISI to create the AI-Medic, a multimodal AI system for autonomous medical mentoring. The system uses a tablet to acquire the view of an operating field. This imagery is given to a neural network trained using DAISI. The network predicts captions describing the steps to be performed in the current view of the surgery. Similar approaches have been used in radiology to describe x-rays and clinical photographs through captions [2]. In our case, however, the model is trained to generate instructions rather than descriptions. Such instructions are conveyed using visual and auditory modalities.

ICMI '20, October 25–29, 2020, Virtual event, Netherlands
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7581-8/20/10.
https://doi.org/10.1145/3382507.3421167

2 METHODOLOGY

Our AI-Medic uses a Deep Learning (DL) model to predict medical instructions from the current view of a surgery. The DL model was training using DAISI. This dataset contains color images and text descriptions of instructions of procedures from various medical disciplines. Using DAISI, an encoder-decoder neural network architecture was trained to predict instructions from input images of medical procedures. The encoder-decoder approach uses a ConvNet as encoder, and a Recursive Neural Network (RNN) as decoder. The ConvNet extracts and encodes visual features from images, and the RNN decodes these visual features into text descriptions.

We use the camera of a tablet device (Galaxy Tab 3, Samsung) to acquire and display the operating field to the user. In addition, every n seconds (as defined by the user), the system will acquire a frame from the live video and analyze whether it presents undesired images artifacts (e.g. blurriness, out-of-focus). If the frame does not present any artifacts, it will be used as input to the DL model. The predicted instructions will be conveyed to the user via two modalities: showing it on the tablet's display, and via text-to-speech.

3 EVALUATING THE AI-MEDIC

We validated our approach using four folds. For each fold, we randomly divided DAISI into train and test sets: approximately 10% of the images were used as test set. The BLEU metric was computed between the predicted and the ground truth instructions to evaluate the algorithm's performance. This is a state-of-the-art metric to evaluate image captioning and machine translation models [3]. BLEU computes a 1-to-100 similarity score by comparing two sentences at the word n-gram level, i.e. analyzing contiguous sequences of n words in a text. We report cumulative BLEU scores for 1-grams to 4-grams for the model's top five candidate predictions.

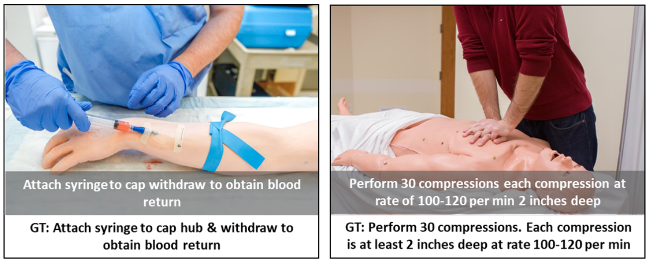
We evaluated the approach by selecting different *Word Count* values: 3, 5, and 7. Only words appearing at least M times during the training phase were part of the model's vocabulary. We also conducted both *Inter-procedure* and *Intra-procedure* evaluations. In *Inter-procedure*, the model had no prior information regarding the procedures in the test set. In *Intra-procedure*, $\frac{1}{P}$ images from each test set procedure were assigned to the training set. In our case, P was set to 0.5. This *Intra-procedure* setting enhanced performance for the test set procedures, otherwise unseen during training.

4 RESULTS

Fig. 2 shows instructions predicted by the AI-Medic. Fig. 3 reports the cumulative BLEU scores for *Inter-procedure* and *Intra-procedure* testing. The captions predicted by our model obtained up to 86 ± 1% 1-gram and 36 ± 1% 4-gram BLEU scores. Overall, the BLEU scores were slightly lower for lower *Word Count* values. A potential reason is that an increased-size vocabulary reduced the chances of learning meaningful relations between the images and the text descriptions.

5 CONCLUSION & FUTURE WORK

This work presented initial steps towards a multimodal AI system for autonomous medical mentoring. The system predicts surgical instructions from input medical images using an encoder-decoder neural network. A tablet device was leveraged to acquire the view of the operating field and convey the predicted instructions to the user via visual and auditory feedback. Future plans to extend the system include adding an attention module to highlight the areas of the operating field that the network is paying attention. This additional modality will be presented into the user's field of view using augmented reality. Overall, this work serves as a baseline for future AI algorithms assisting in autonomous medical mentoring.



Attach syringeto cap withdraw to obtain blood return

GT: Attach syringe to cap hub & withdraw to obtain blood return

Perform 30 compressions each compression at rate of 100-120 per min 2 inches deep

GT: Perform 30 compressions. Each compression is at least 2 inches deep at rate 100-120 per min

Figure 2: Examples of instructions predicted by the AI-Medic. The predicted instruction is in white font, inside the images. The ground truth (GT) instruction is written below.

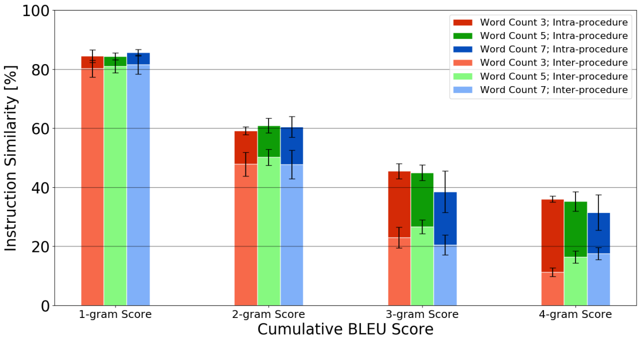


Figure 3: Cumulative *n*-gram BLEU scores. Our model was evaluated using three *Word Count* values (3, 5, 7) and two testing approaches (*Inter-procedure, Intra-procedure*).

ACKNOWLEDGMENTS

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-14-1-0042. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the funders.

REFERENCES

[1] Magdala de Araújo Novaes and Arindam Basu. 2020. Disruptive technologies: Present and future. In *Fundamentals of Telemedicine and Telehealth*. Elsevier, 305–330.

[2] Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. 2019. A Survey on Biomedical Image Captioning. *arXiv preprint arXiv:1905.13302* (2019).

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[4] Edgar Rojas-Muñoz, Kyle Couperus, and Juan Wachs. 2020. DAISI: Database for AI Surgical Instruction. *arXiv preprint arXiv:2004.02809* (2020).

[5] Edgar Rojas-Muñoz, Chengyuan Lin, Natalia Sanchez-Tamayo, Maria Eugenia Cabrera, Daniel Andersen, Voicu Popescu, Juan Antonio Barragan, Ben Zarzaur, Patrick Murphy, Kathryn Anderson, et al. 2020. Evaluation of an augmented reality platform for austere surgical telementoring: a randomized controlled crossover study in cricothyroidotomies. *NPJ Digital Medicine* 3, 1 (2020), 1–9.