RESEARCH ARTICLE

Spatio-temporal deep learning model for distortion classification in laparoscopic video [version 1; peer review: awaiting peer review]

Nouar Aldahoul ^{1,2}, Hezerul Abdul Karim ¹, Abdulaziz Saleh Ba Wazir ¹, Myles Joshua Toledo Tan^{2,3}, Mohammad Faizal Ahmad Fauzi ¹

¹Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

²YO-VIVO corporation, Bacolod City, 6100, Philippines

³Department of Natural Sciences, University of St. La Salle, Bacolod City, 6100, Philippines

V1 First published: 05 Oct 2021, 10:1010
<https://doi.org/10.12688/f1000research.72980.1>
Latest published: 05 Oct 2021, 10:1010
<https://doi.org/10.12688/f1000research.72980.1>

Open Peer Review

Approval Status Awaiting Peer Review

Any reports and responses or comments on the article can be found at the end of the article.

Abstract

Background: Laparoscopy is a surgery performed in the abdomen without making large incisions in the skin and with the aid of a video camera, resulting in laparoscopic videos. The laparoscopic video is prone to various distortions such as noise, smoke, uneven illumination, defocus blur, and motion blur. One of the main components in the feedback loop of video enhancement systems is distortion identification, which automatically classifies the distortions affecting the videos and selects the video enhancement algorithm accordingly. This paper aims to address the laparoscopic video distortion identification problem by developing fast and accurate multi-label distortion classification using a deep learning model. Current deep learning solutions based on convolutional neural networks (CNNs) can address laparoscopic video distortion classification, but they learn only spatial information.

Methods: In this paper, utilization of both spatial and temporal features in a CNN-long short-term memory (CNN-LSTM) model is proposed as a novel solution to enhance the classification. First, pre-trained ResNet50 CNN was used to extract spatial features from each video frame by transferring representation from large-scale natural images to laparoscopic images. Next, LSTM was utilized to consider the temporal relation between the features extracted from the laparoscopic video frames to produce multi-label categories. A novel laparoscopic video dataset proposed in the ICIP2020 challenge was used for training and evaluation of the proposed method.

Results: The experiments conducted show that the proposed CNN-LSTM outperforms the existing solutions in terms of accuracy (85%), and F1-score (94.2%). Additionally, the proposed distortion identification model is able to run in real-time with low inference time (0.15 sec).

Conclusions: The proposed CNN-LSTM model is a feasible solution to be utilized in laparoscopic videos for distortion identification.

Keywords

distortion classification, convolutional Neural Network, laparoscopic video, long short-term memory, multi-label classification, spatio-temporal features

This article is included in the **Artificial Intelligence and Machine Learning** gateway.

This article is included in the **Research Synergy Foundation** gateway.

Corresponding author: Nouar AIDahoul (nouar.aldahoul@live.iium.edu.my)

Author roles: **AIDahoul N:** Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Abdul Karim H:** Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Ba Wazir AS:** Methodology, Writing – Original Draft Preparation; **Toledo Tan MJ:** Formal Analysis, Validation, Writing – Review & Editing; **Ahmad Fauzi MF:** Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research project was funded by Multimedia University, Malaysia.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 AIDahoul N *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: AIDahoul N, Abdul Karim H, Ba Wazir AS *et al.* **Spatio-temporal deep learning model for distortion classification in laparoscopic video [version 1; peer review: awaiting peer review]** F1000Research 2021, **10**:1010 <https://doi.org/10.12688/f1000research.72980.1>

First published: 05 Oct 2021, **10**:1010 <https://doi.org/10.12688/f1000research.72980.1>

Introduction

Video quality assessment (VQA) in the medical field is an important task to achieve satisfactory conditions for medical imaging modalities like magnetic resonance imaging (MRI), computed tomography (CT) scans, and laparoscopy. VQA is composed of two stages: distortion classification and quality score evaluation. Laparoscopic surgery videos are prone to distortions that affect a surgeon's visibility and degrade the vision quality for robot-assisted surgery.¹

Laparoscopic videos are often affected by various types of distortions like noise, smoke, uneven illumination, and blur, which are all concomitant artifacts that arise from operating the laparoscopic surgical equipment.² To enhance the distorted laparoscopic videos, most studies propose solutions that require troubleshooting the equipment.^{2,3} However, such solutions are time consuming and cannot guarantee high-quality laparoscopy every time.

Recent studies have suggested the use of image or video enhancement methods like de-smoking for laparoscopic surgery,⁴⁻⁶ and joint wavelet decomposition and binocular combination for endoscopic image enhancement.⁷ In this case, real-time detection of the types of distortion is important to decide which enhancement methods are appropriate to apply. Real-time distortion classification is a challenging task and few recent studies have addressed it using hand-crafted features.⁸⁻¹² These existing image quality assessment methods, such as BIQL,¹¹ DIIVINE¹² and BRISQUE,¹⁰ were based on non-generic classification and are considered domain-dependent tasks. In addition, a distortion-specific classification approach has been demonstrated.⁸ This approach used a separate traditional feature method for each type of distortion.⁸

On the other hand, convolutional neural networks (CNNs) overcome the previous limitations and learn features automatically with the same CNN architecture to detect all types of distortions. This paper aims to address the challenge of distortion detection and produce a generic method for distortion classification in laparoscopic videos.

Artificial neural networks (ANNs) have shown significant capability in overcoming the issue of distortion classification by extracting informative features from all kinds of distortions. CNNs are powerful and efficient in several image tasks including classification,¹³ segmentation,¹⁴ enhancement,¹⁵ and retrieval.¹⁶ Recently, CNNs have also been used in several studies on image distortion classification for various applications.^{17,18} However, recurrent neural networks (RNNs), and specifically, long short-term memory (LSTM)¹⁹ have not yet been investigated for distortion classification in video datasets. This paper aims to highlight the use of CNN-LSTM²⁰ to improve classification accuracy.

In the context of distortion classification in laparoscopic surgery videos, a recent study has proposed the use of deep CNNs, such as ResNet for distortion ranking.²¹ Its method achieved ranking accuracies of 83.3%, 84.7%, and 87.3% using Resnet18, Resnet34, and Resnet50, respectively. However, the previous work focused only on spatial features extracted from a collection of 20,000 images for image-level distortion ranking.

Another very recent work was found to transfer learning from pre-trained ResNet50 CNN to laparoscopic video frames.²² The spatial features extracted from ResNet50 were applied to four support vector machine classifiers (three binary and one 5-class) utilizing decision fusion to produce the final distortion lists.²² Hence, this paper proposes to extract spatiotemporal features using CNN-LSTM for video-level distortion classification.

The key contributions of this paper are:

- Utilization of a RNN model such as LSTM with time series of CNN-based features extracted from the frames. To the best of our knowledge, this is the first paper that uses CNN-LSTM for non-reference distortion classification in laparoscopic videos.
- An evaluation and comparison between the proposed CNN-LSTM and existing solutions presented for the ICIP2020 challenge.

This paper is structured as follows: *Methods* describes the proposed method and the experiments including the dataset and the experimental setup. In *Results and discussion*, the results of the proposed solution and the comparison with existing methods are presented and discussed. *Conclusions* summarizes the significance of this work and opens doors for further improvement.

Methods

The proposed multi-label distortion classification with CNN-LSTM

In this section, we describe the proposed methodology for distortion classification in laparoscopic videos. This classification problem is formulated as a single multi-label classification which can be transformed to multiple binary

Figure 1. Illustration of the proposed multi-label distortion classification. CNN, convolutional neural networks; LSTM, long short-term memory; AWGN, additive white Gaussian noise.

classifiers. In this scenario, each label (distortion) in the dataset is used with a separate binary classifier, resulting in five binary classifiers in total. The block diagram of the proposed model is shown in Figure 1.

Transfer learning with residual network

Usually, very deep CNNs suffer from the gradient vanishing problem, which leads to a drop in accuracy.²³ To address this problem, residual network (ResNet) was developed utilizing skip connections instead of direct stacked layers.²³ ResNet is a well-known deep neural network with high generalization ability used for image recognition.²³ Residual networks have various versions with different numbers of layers, such as ResNet50 with 50 layers and over 23 million trainable parameters.

The transfer learning approach is summarized by training deep CNNs like ResNet with a large-scale dataset such as ImageNet²⁴ and utilizing them with a novel small-scale dataset. In this paper, ResNet50²³ was transferred to the laparoscopic video dataset and utilized to extract spatial features from the video's frames. This CNN pre-trained on ImageNet²⁴ was used after removing top layers. The input images were resized to 224×224 and the dimensions of extracted features was 2048.

Classification with LSTM

LSTM is a special type of RNN that is used for long-range sequence modeling.¹⁹ LSTM has a memory cell, which acts as an accumulator of state information, supported by control gates. The advantage of this structure is that it solves the problem of gradient vanishing.¹⁹ The CNN-LSTM network was found to capture spatiotemporal correlations better than fully connected LSTM, which is only powerful for spatial correlation.²⁰

In this paper, the spatial feature vector extracted from ResNet50 represents one laparoscopic frame. Additionally, the series of feature vectors extracted from a series of frames in one video was applied to a set of five LSTMs. This aims to map the video to two categories in each LSTM. For example, the first LSTM checks whether smoke distortion is available in a video and produces two classes: "yes" and "no." The already-trained CNN was utilized after replacing the top layers with five LSTM classifiers to tune the parameters of the fully connected layers. In other words, each LSTM fits the extracted features and maps them to two categories: "yes" and "no."

The architecture of each LSTM consists of the following layers:

- 1) Bidirectional LSTM with 64 nodes
- 2) ReLU activation function
- 3) Batch normalization
- 4) Dropout with 0.2

- 5) Fully connected layers with 64 nodes
- 6) ReLU activation function
- 7) Batch normalization
- 8) Dropout with 0.2
- 9) Fully connected layers with two nodes
- 10) Softmax activation function

Experiments

Datasets and experimental setup

The dataset used in this paper is an extended version of the Laparoscopic Video Quality (LVQ) database.⁸ The database contains 10 reference videos, each 10 seconds in length.⁸ Each reference video is distorted by five different types of distortions with four different levels, resulting in a total of 200 videos. These videos were extracted from the Cholec80 dataset that comprises 80 different videos of cholecystectomy surgeries.²⁵ The extracted videos were selected considering multiple variations of scene content. The resolution of the videos is 512×288 with a 16:9 aspect ratio and a frame rate of 25 fps.

The extended version of LVQ dataset was issued in the ICIP2020 challenge and includes 1000 laparoscopic videos divided into 800 videos for training and 200 videos for testing. The distortions include additive white Gaussian noise (AWGN), smoke, uneven illumination, defocus and motion blur. The numbers of videos for each label or distortion are not balanced (300 videos with AWGN, 320 videos with smoke, 400 videos with uneven illumination, 160 videos with defocus blur, 80 videos with motion blur). The challenge in this dataset is that each video is affected by single or multiple distortions and thus, the problem of distortion classification is formulated as a multi-label classification problem.

The training and testing for the ResNet-LSTM model was carried out using **OpenCV** and **TensorFlow** frameworks and libraries on an NVIDIA GeForce GTX 1080 Ti GPU. The learning rate used to train the LSTM model was set to 0.001, the batch size was set to 8, and the number of epochs was set to 150. The minimization of the categorical crossentropy loss function was achieved using the Adam optimizer.

Results and discussion

To the best of our knowledge, no other papers have utilized this extended version of the laparoscopic video dataset **challenge dashboard** for distortion classification. For this reason, we compared our approach with the best solutions presented in the ICIP2020 challenge as shown in **Table 1**.

The description of the baseline solutions was given by winners in the **ICIP2020 challenge presentation event**. One of the solutions was based on using a VGG16 CNN²⁶ to extract features. The feature vector was applied to the fully connected

Table 1. Classification accuracy and F1-score of the proposed method and various baseline models.

Solution	F1-score (single + multi distortions)	F1-score (single-distortion)	Accuracy
VGG16 + many fm + fc (Baseline)*#	94.1%	93.3%	81.5%
VGG16 + 5 fc (Baseline)*#	93.3%	90.7%	78.0%
(Baseline)*	91.5%	88.0%	76.5%
(Baseline)*	85.4%	98.7%	58.0%
(Baseline)*	83.2%	89.3%	57.0%
ResNet50-LSTM (Proposed)	94.2%	89.3%	85.0%

Data sources: * = **Challenge dashboard**; # = **Challenge presentation event**.

neural network that included two hidden layers with 4096 nodes, two batch normalization layers, and two dropout layers. On the other hand, another solution used a deep multi-task learning model. It included one shared VGG-based feature extraction block and five independent binary classifiers (one for each distortion type). Each classifier had two fully connected layers with 512 nodes and one node in the output layer with a sigmoid activation function [ICIP2020 challenge presentation event](#). The description of other baseline solutions was not presented, but the results were shown in the [challenge dashboard](#).

The performance of the proposed methodology was evaluated in terms of classification accuracy, F1-score of single distortion, and F1-score of single and multiple distortions as shown in [Table 1](#). It can be observed that the proposed ResNet50-LSTM leads to the best accuracy of 85.0%, while baseline methods yielded accuracies of between 57% and 81.5%. Additionally, ResNet50-LSTM yielded the best F1-score of single and multiple distortions (94.2%), while baseline methods yielded F1-score between 83.2% and 94.1%. Furthermore, it is clear that the performance of our method for multiple distortions outperforms that for single distortion, which still has room for improvement.

[Figure 2](#) shows the confusion matrix for each distortion category produced from each LSTM. The LSTMs were able to correctly classify 58 videos out of 60, 46 videos out of 50, 94 out of 95, 88 out of 95 for AWGN, defocus blur, smoke, and uneven illumination, respectively. On the other hand, motion blur LSTM gave the worst classification performance with 29 correct videos out of 45. The reason for this drop was that the videos with motion blur have the minimum number of samples, which is only 80 videos. The performance of the motion LSTM can be improved significantly by having more samples affected by motion blur distortion. The performance metrics of the proposed method for each class are shown in [Table 2](#).

The proposed ResNet50-LSTM was able to run considering real-time conditions. The inference time was 0.05 seconds to extract features from one frame using ResNet50. The features extracted from one frame were added to the features of other frames to be applied to the LSTM. The inference time for five LSTMs to produce the five distortion classes was 0.1 seconds. In summary, the proposed model updates the distortion categories every 0.15 seconds and achieves high speed performance.

Figure 2. Confusion matrix of a) additive white Gaussian noise, b) defocus blur, c) motion blur, d) smoke, e) uneven illumination.

Table 2. Performance metrics of the proposed method for each class in the laparoscopic dataset.

Distortion	Accuracy %	Recall %	Precision %	F1-score %	FNR %	FPR %
AWGN noise	97.5	96.66	95.08	95.86	3.33	2.14
Defocus blur	97.0	92.0	95.83	93.88	8.0	1.33
Motion blur	91.0	64.44	93.55	76.31	35.56	1.29
Smoke	98.5	98.95	97.92	98.43	1.05	1.90
Uneven illumination	96.5	92.63	100	96.17	7.37	0
Average	96.1	88.94	96.48	92.13	11.06	1.33

AWGN, additive white Gaussian noise; FNR, false negative rate; FPR, false positive rate.

Conclusions

In this paper, a novel strategy of distortion classification was proposed. A multi-label spatiotemporal deep model, including a pre-trained deep CNN of ResNet50 and five LSTMs, was used to address the problem of single and multiple distortion classification. The proposed model was tested with a laparoscopic video dataset and the results were promising. It was found that our model outperformed existing solutions in terms of accuracy by 4.5% and yielded the best F1-score for single and multiple distortions. Hence, we intend to enhance the performance by tuning more layers of pre-trained CNN with laparoscopic images affected by distortions to learn more informative features. The last step requires collecting a large number of images to achieve promising improvements. Additionally, more recent CNN architectures such as EfficientNet²⁷ and DeiT (Data-efficient Image Transformers)²⁸ models are good candidates for extracting informative features. In this paper, the proposed solution only classifies the laparoscopic distortions into five categories. Hence, in future work, we plan to rank each category of distortion in terms of distortion intensity, which is a more challenging matter.

Data availability

Underlying data

The datasets used in this work were used for the ICIP 2020 challenge and created by researchers from Université Sorbonne Paris Nord, France; Norwegian University of Science and Technology, Norway; and Oslo University Hospital, Norway. The datasets are publicly available under a CC-BY-NC-SA 4.0 license from <https://github.com/zakopz/icip-2020-lvq-challenge>.

This dataset was not generated nor is it owned by the authors of this article; the listed owners are Université Sorbonne Paris Nord, France; Norwegian University of Science and Technology, Norway; and Oslo University Hospital, Norway. Therefore, neither the authors nor F1000Research are responsible for the content of this dataset and cannot provide information about data collection. As this dataset contains potentially identifying images/information, caution is advised when using this dataset in future research.

References

- Sánchez-González P, et al.: **Laparoscopic video analysis for training and image-guided surgery.** *Minim Invasive Ther Allied Technol.* 2011; **20**(6): 311–320.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Verdaasdonk EGG, Stassen LPS, van der Elst M, et al.: **Problems with technical equipment during laparoscopic surgery: An observational study.** *Surg Endosc.* 2007; **21**(2): 275–279.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Siddaiah-Subramanya M, Nyandowe M, Tiang KW: **Technical problems during laparoscopy: A systematic method of troubleshooting for surgeons.** *Innov Surg Sci.* 2017; **2**(4): 233–237.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang C, Cheikh FA, Kaaniche M, et al.: **A smoke removal method for laparoscopic images.** *arXiv.* 2018: 6–10.
- Venkatesh V, Sharma N, Srivastava V, et al.: **Unsupervised smoke to desmoked laparoscopic surgery images using contrast driven**
Cyclic-DesmokeGAN. *Comput Biol Med.* 2020; **123**: 103873.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wang C, Mohammed AK, Cheikh FA, et al.: **Multiscale deep desmoking for laparoscopic surgery.** *SPIE Medical Imaging 2019.* 2019, p. 68.
[Publisher Full Text](#)
- Sdiri B, Kaaniche M, Cheikh FA, et al.: **Efficient enhancement of stereo endoscopic images based on joint wavelet decomposition and binocular combination.** *IEEE Transactions Medical Imaging.* 2019; **38**(1): 33–45.
[Publisher Full Text](#)
- Khan ZA, et al.: **Towards a video quality assessment based framework for enhancement of laparoscopic videos.** *arXiv.* 2020.
[Publisher Full Text](#)
- Khan ZA, Kaaniche M, Beghdadi A, et al.: **Joint Statistical Models for No-Reference Stereoscopic Image Quality Assessment.** *Proc Euro Workshop Visual Information Processing, EUVIP.* 2018, pp. 26–28.
[Publisher Full Text](#)

10. Mittal A, Moorthy AK, Bovik AC: **No-reference image quality assessment in the spatial domain.** *IEEE Trans. Image Process.* 2012; **21**(12): 4695–4708.
[Publisher Full Text](#)
11. Moorthy AK, Bovik AC: **A two-stage framework for blind image quality assessment.** *IEEE Signal Processing Letters.* 2010; **17**(5): 513–516.
[Publisher Full Text](#)
12. Moorthy AK, Bovik AC: **Blind image quality assessment: From natural scene statistics to perceptual quality.** *Trans. Image Process.* 2011; **20**(12): 3350–3364.
[Publisher Full Text](#)
13. Jiang J, Feng X, Liu F, et al.: **Multi-Spectral RGB-NIR Image Classification Using Double-Channel CNN.** *IEEE Access.* 2019; **7**: 20607–20613.
[Publisher Full Text](#)
14. Kim Y, Kim S, Kim T, et al.: **CNN-based semantic segmentation using level set loss.** *2019 IEEE Winter Conference on Applications of Computer Vision, WACV.* 2019, **2019**, pp. 1752–1760.
[Publisher Full Text](#)
15. Qiu T, Wen C, Xie K, et al.: **Efficient medical image enhancement based on CNN-FBB model.** *IET Image Processing.* 2019; **13**(10): 1736–1744.
[Publisher Full Text](#)
16. Alex V, Khened M, Ayyachamy S, et al.: **Medical image retrieval using Resnet-18 for clinical diagnosis.** *SPIE Medical Imaging.* 2019; **1095410**: 35.
[Publisher Full Text](#)
17. Hossain MT, Teng SW, Zhang D, et al.: **Distortion Robust Image Classification Using Deep Convolutional Neural Network with Discrete Cosine Transform.** *Int Conf Image Processing (ICIP).* 2019, pp. 659–663.
[Publisher Full Text](#)
18. Buczkowski M, Stasinski R: **Convolutional Neural Network-Based Image Distortion Classification.** *2019 Int Conf Systems Signals Image Processing (IWSSIP).* 2019, pp. 275–279.
[Publisher Full Text](#)
19. Hochreiter S, Schmidhuber J: **Long Short-Term Memory Neural Comput.** 1997.
[Publisher Full Text](#)
20. Islam MZ, Islam MM, Asraf A: **A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images.** *Informat. Med. Unlocked.* 2020; **20**.
[Publisher Full Text](#)
21. Khan ZA, Beghdadi A, Kaaniche M, et al.: **Residual Networks Based Distortion Classification and Ranking for Laparoscopic Image Quality Assesment.** *2020 IEEE Int Conf Image Processing (ICIP).* 2020, pp. 176–180.
[Publisher Full Text](#)
22. Aldahoul N, Karim HA, Tan MJT, et al.: **Transfer Learning and Decision Fusion for Real Time Distortion Classification in Laparoscopic Videos.** *IEEE Access.* 2021; **9**: 115006–115018.
[Publisher Full Text](#)
23. He K, Zhang X, Ren S, et al.: **Deep residual learning for image recognition.** 2016.
[Publisher Full Text](#)
24. Deng J, Dong W, Socher R, et al.: **ImageNet: A large-scale hierarchical image database.** 2010.
[Publisher Full Text](#)
25. Twinanda AP, Shehata S, Mutter D, et al.: **EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos.** *IEEE Trans. Med. Imaging.* 2017.
[Publisher Full Text](#)
26. Simonyan K, Zisserman A: **Very deep convolutional networks for large-scale image recognition.** 2015. arXiv: 1409.1556.
27. Tan M, Le QV: **EfficientNet: Rethinking model scaling for convolutional neural networks.** 2019. arXiv:1905.11946.
28. Touvron H, Cord M, Douze M, et al.: **Training data-efficient image transformers& distillation through attention.** 2020; arXiv: 2012.12877.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research