DEEP SMOKE REMOVAL FROM MINIMALLY INVASIVE SURGERY VIDEOS

Sabri Bolkar[*]*, Congcong Wang*[†]*, Faouzi Alaya Cheikh, Sule Yildirim*

Norwegian Colour and Visual Computing Laboratory, NTNU, Norway

ABSTRACT

During video-guided minimally invasive surgery, quality of frames may be degraded severely by cauterization-induced smoke and condensation of vapor. This degradation of quality creates discomfort for the operating surgeon, and causes serious problems for automatic follow-up processes such as registration, segmentation and tracking. This paper proposes a novel deep neural network based smoke removal solution that is able to enhance the quality of surgery video frames in real-time. It employs synthetically generated training dataset including smoke embedded and clean reference versions. Results calculated on the test set indicate that our network outperforms previous defogging methods in terms of quantitative and qualitative measures. While eliminating apparent smoke, it also successfully preserves the natural appearance of tissue surface. To the best of our knowledge, the presented method is the first deep neural network based approach for the surgical field smoke removal problem.

Index Terms— Image restoration, smoke removal, defogging, deep image processing, convolutional neural networks

1. INTRODUCTION

In laparoscopic surgery (i.e., minimally invasive surgery), operations are performed through small incisions where instruments such as camera and dissection tools are introduced to the body [1]. In modern laparoscopy, the camera is utilized as the main observation unit, and a video processing pipeline accompanying mono/stereo camera is becoming widespread as it allows for segmentation, registration and image-based navigation during surgery [2].

Although laparoscopic surgery is more comfortable operation for the patient compared to open surgery, it brings several important challenges. Video frames deteriorate due to severe visibility loss occurring because of smoke induced by tissue dissection tools (e.g., electrocautery, laser tissue ablation and ultrasonic scalpel) and vapor condensation resulting from temperature difference between body and the operation room [3]. Loss of visibility makes removal of the smoke and cleaning of the camera lenses a critical issue during the operation.

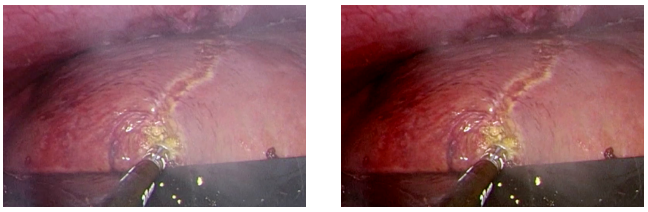

Fig. 1: Smoke removal example applied on a sample laparoscopic surgery video frame. Left: An example frame with smoke. Right: Desmoked by our method.

Although there are several chemical and mechanical solutions proposed by biomedical industry (e.g., Pall Laparoshield™ smoke filtration system), an automatic and real-time alternative is highly desired.

Inspired by Li *et al.* [4], we utilize convolutional architecture with multi-scale kernels and synthetically rendered smoke dataset in our method. The trained network accompanied with a demo code will be publicly available at github.com/elras/desmokenet. The main contributions of this work are as following:

1. This work illustrates the first known application of convolutional neural networks to the surgical smoke removal problem.
2. We are the first to employ synthetic smoke for generating training dataset including ground truth clean tissue images and smoke embedded versions for surgery field desmoking.
3. As it will be shown in the results section, the trained network performs quite well and obtains the best results in terms of full reference MSE, PSNR and MAD [5] metrics compared to previous defogging methods [4, 6, 7, 8].
4. The proposed method preserves the natural tissue color without needing any augmentation to the network. This end-to-end approach reaches processing speed up to $\sim$ 20 fps on a single GPU.

The remainder of the paper is organized as follows. In Sec. 2, we review the existing desmoking techniques. The synthetic dataset and network model are presented in Sec. 3. Sec. 4 presents training of the network and the experimental results. Conclusions are drawn in Sec. 5.

[*]Sabri Bolkar is supported by COSI Erasmus Mundus Master Scholarship

[†]Congcong Wang is funded by the Research Council of Norway through project no. 247689 IQ-MED.

2. RELATED WORK

Although dehazing research is well established, smoke removal from surgery videos is a recent topic of interest and there exist only four known published studies [9, 10, 11, 12]. All of these works share a common framework as they first apply one of the already existing dehazing solutions and augment the method to decrease observed deterioration of the natural tissue appearance. In the literature, hazing is represented by atmospheric scattering equation, and the relation of the hazed image $I(x)$ to transmission map $t(x)$ and haze-free image $J(x)$ is expressed in Eq. (1) where spatial location is denoted by x. Using this hazing equation, dehazing process reduces to recovery of $J(x)$ by estimating $t(x)$ and atmospheric light A [13].

$$I(x) = t(x)J(x) + A(1 - t(x)). \quad (1)$$

The transmission map $t(x)$ for each pixel x can be defined as Eq. (2), where $d(x)$ is the depth map and β is the scattering coefficient [13]:

$$t(x) = e^{-\beta d(x)}. \quad (2)$$

Previous works can be divided into three main categories as we will discuss in the sections below: Desmoking using refined dark channel prior [9], desmoking using Bayesian inference [10, 11] and desmoking using visibility-driven fusion [12].

Desmoking using refined dark channel prior. Tchaka *et al.* [9] propose to use dark channel prior dehazing method (DCP) [6] as baseline and refine the pipeline presented in the original work. Authors propose two heuristic improvements to prevent the color distortion when DCP is directly applied: Thresholding the dark channel by a constant value to eliminate outliers and de-emphasizing pixel values in a certain range where smoke is expected to be present. To further improve color and contrast, histogram equalization is applied as the last step in the pipeline. To compare the performance of their method, authors take a single frame without smoke at the outset of a test video as the reference, and compute mean error between the reference frame and following several frames with smoke. Their method shows decrease in mean error for the processed test images when compared with the unprocessed images.

Desmoking using Bayesian inference. Inspired by Nishino *et al.* [14], Kotwal *et al.* [10] and Baild *et al.* [11] offer Bayesian inference-based laparoscopy image desmoking accompanied with denoising and specularity removal in addition to denoising, respectively. Authors represent the uncorrupted image as a Markov random field and apply maximum a posteriori estimation to obtain enhanced versions. Unlike our method, realistic smoke generation is not carried out.

Desmoking using visibility-driven fusion. Luo *et al.* [12], inspired by [15], propose Poisson fusion based defogging method in addition to reformulation of atmospheric scattering equation, Eq. (1), which is claimed to decrease computational load. Evaluation of the method is carried out by both subjective and objective measures. Authors collect images during robotic-assisted laparoscopic radical prostatectomy surgery to compare their results with previously proposed defogging methods. Because of lacking reference images, no-reference based naturalness [16] and sharpness [17] metrics are used. The results show slightly better naturalness and equivalently good sharpness to state of the art.

3. DATASET AND DEEP NETWORK MODEL

Training of deep neural networks require big labeled datasets. In dehazing, since the image haze is dependent on depth, synthetic haze dataset generation using available depth map datasets is widely employed (e.g., NYU depth dataset [18]). However, there exists no available dataset that can be used for training of a desmoking network. The unique appearance of tissues also prevents us from employing available natural image datasets. In this paper, we propose to utilize computer graphics to generate synthetic smoke and embed the generated smoke to clean video frames. To collect ground truth smoke-free images, Hamlyn Centre Laparoscopic and Endoscopic Dataset videos are chosen [19, 20]. From stereo videos, only left camera frames are extracted. Also to eliminate present natural smoke, each frame is manually checked and frames with natural smoke are discarded from the set. In total, approximately 19.600 images are collected. From this collection, random 100 images are selected to be used as the test set.

3.1. Synthetic Smoke Generation and Embedding

Perlin noise is a well-established method that has been employed by computer graphics community for years to create natural appearance of rough surfaces such as terrain and mountain. It can also be used to render artificial fire, cloud and smoke images [21]. It is relatively straightforward to implement and quite flexible as parameters can be modified to change the roughness of the surface. Perlin noise is generated at each pixel in the image space by computing a pseudo-random gradient at each of the eight nearest vertices on the integer cubic lattice, then cubic splined interpolation is applied to find the desired pixel value [22]. To enable realistic rendering, noises with different frequency setting are added as well. In our method, we exploit Perlin noise to generate synthetic smoke. To embed the generated smoke images, we take an heuristic approach and utilize linear mixture:

$$I_e^c(x) = I_g^c(x) + 0.8(I_s^c(x) - 1/N \sum_{i=1}^{N} I_s^c(i)), \quad (3)$$

where I_e^c is the smoke embedded image, I_g^c is the clean ground truth image, I_s^c is the generated Perlin smoke and N

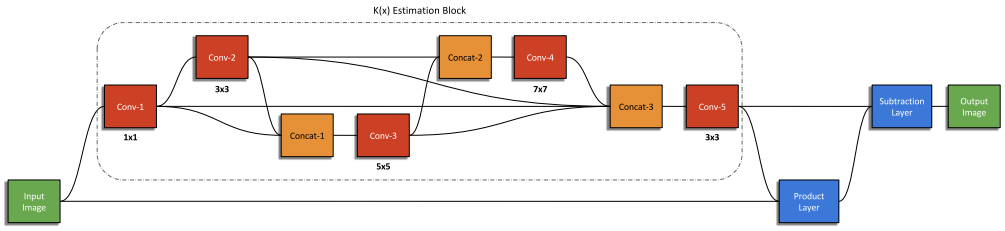

Fig. 2: Schematic illustrating the network architecture. The network includes five convolutional and three concatenating layers which take inputs from feature vectors with different scales. Output image is estimated directly from the input image via Eq. (4).

is the number of pixels for each color channel c $\in$ {R, G, B}. The coefficient 0.8 is chosen according to our experiments to ensure realistic appearance.

Before training, preprocessing is applied to all images in the dataset. Firstly, images are cropped such that corrupted 10 pixel wide frame from all sides is removed, subsequently images are also heuristically resized to 512x512. It is known that scaling and subtracting mean improves convergence of the network, but we decided to keep the mean and only rescaled the images to [0, 1] interval. The reason for this is the desire to keep the natural color of the tissues. Higher average pixel value observed in smoked images is an important cue that can be used to improve the network's performance.

3.2. Desmoking Deep Network Model

At the outset, we had several attempts to employ a novel convolutional model. However, the dataset includes statistically similar images retrieved from surgery videos and trained networks performed poorly. Lack of available surgery datasets also restricts development of a new model, thereby motivating us to employ transfer learning by initializing filter weights from a trained dehazing network.

Similarity between haze and smoke is apparent as both deteriorate visibility by decreasing contrast, saturation and increasing overall lightness. Unlike haze however, physics governing smoke appearance cannot be simply represented by depth-dependent atmospheric scattering equation. Smoke appearance is more heterogeneous and in our case observed from close range. To be able to employ transfer learning, we require an intermediate parameter that is flexible enough to enable learning of weights for surgery smoke removal. To date, there exist three known networks proposed for dehazing. The first is DehazeNet [7], Ren *et al*. propose a multiscale neural network approach [23] and the more recent one inspired by Ren *et al.* is All-in-One Dehazing Network (AOD-Net) [4].

Although these networks are specifically designed for natural scene dehazing, AOD-Net reformulates the hazing equation. Rather than calculating haze-free image by first estimating the transmission map, it aims to output the haze-free output image directly by jointly estimating A and $t(x)$ in a parameter denoted $K(x)$, hence allowing deep filters to learn internal dehazing features [4]. AOD-Net model incorporates $K(x)$ in Eq. (5) into the scattering equation Eq. (1) to retrieve haze-free image $J(x)$ as in Eq. (4) (where $b = 1, \forall$ x):

$$J(x) = K(x)I(x) - K(x) + b, \tag{4}$$

$$K(x) = \frac{\frac{1}{t(x)}(I(x) - A) + (A - b)}{I(x) - 1}. \tag{5}$$

This formulation of the equation allows the neural network to estimate image-dependent $K(x)$ using the hazed input image itself. Later in the pipeline, $J(x)$ is calculated by pixel-wise linear operations, in Eq. (4), thus favoring processing speed.

In this study, we propose to use AOD-Net model for transfer learning of smoke removal as shown in Fig. 2. The net consists of five convolutional layers with ReLU activation units and three concatenating layers where features from multiscale kernels are combined. Filter size of convolutional layers are unity for *Conv-1*, 3 for *Conv-2* and *Conv-5*, 5 for *Conv-3* and 7 for *Conv-4*. Output of the convolutional layers at different levels are combined such that *Concat-1* takes input from *Conv-1* and *Conv-2*, *Concat-2* concatenates *Conv-2* and *Conv-3* features, lastly *Concat-3* takes from *Conv-1*, *Conv-2*, *Conv-3* and *Conv-4*. The last two layers apply pixel-wise multiplication and subtraction operations to obtain smoke-free image.

4. TRAINING AND RESULTS

Training is carried out by using Caffe deep learning framework [24]. We fine-tune all layers by initializing the network weights from original AOD-Net weights. The training batch size for stochastic gradient descent is chosen to be 8, input and output image has size of 512x512 in sRGB color space. As the loss function, Mean Squared Error (MSE) is employed. Initial learning rate, momentum and decay rate is selected to be 0.0001, 0.9 and 0.00001 for all layers. Learning rate is halved at 2nd, 8th and 12th epochs. Optimum performance for the network is reached after 16 epochs. It is also seen that higher learning rates decrease the quality of the results in fine-tuning stage. We also tried to vary learning rates across layers, but the network performed without any noticeable improvement.

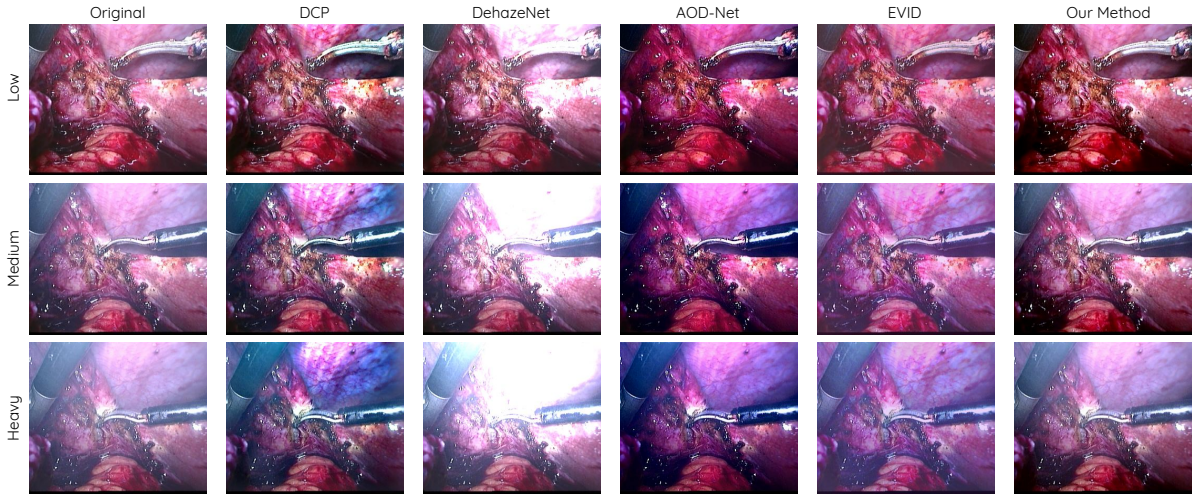

Fig. 3: Qualitative comparison of methods with three different smoke density levels. Our method shows minimum distortion on appearance of the tissue surface when compared to DCP, Dehazenet, AOD-Net and EVID.

4.1. Results and Discussion

To compare the performance of our method with state of the art, both quantitative and qualitative evaluations are performed. For quantitative evaluation, we employed the reserved test set which includes 100 samples with ground truth and computed MSE, PSNR and MAD (Most-Apparent-Distortion) [5] by considering Pedersen's work [25]. Because of lacking availability of published codes described in Sec. 2, three methods based on atmospheric model, Dark Channel Prior (DCP) [6], DehazeNet [7], AOD-Net [4] and variational histogram optimization based EVID [8] are compared with our network. For qualitative evaluation, a surgery video with induced smoke from Hamlyn dataset [26] is utilized.

Method	Full Reference Evaluation Results					
	Avr. MSE	Std. MSE	Avr. PSNR	Std. PSNR	Avr. MAD	Std. MAD
DCP [6]	1.51	0.89	18.59	3.28	116.48	14.76
DehazeNet [7]	3.09	1.77	15.36	2.85	125.07	8.82
AOD-Net [4]	1.42	0.40	18.36	2.11	118.52	**7.26**
EVID [8]	1.07	0.46	19.45	1.66	117.20	7.68
Our Method	**1.00**	**0.36**	**19.72**	**1.57**	**97.85**	8.66

Table 1: Average and standard deviation results for each evaluation metric. MSE values are normalized with respect to the average MSE of our method. Bold values indicate superior performance.

As tabulated in table 1, our method outperforms the previous defogging methods in terms of MSE, PSNR and especially MAD with a high margin. Observed small standard deviation is an indicative of high robustness of our network. Fig. 3 presents the results of methods for three frames with minimum (first row), medium (second row) and heavy smoke (last row) with reference frames in the first column. DCP alters the color of the surface and results in an unnatural color appearance, DehazeNet outputs heavily saturated images, AOD-Net results in over-enhanced surfaces with purplish color and EVID similarly corrupts the natural tissue appearance especially for the heavy smoke case. It is interesting to note that although our network borrows the AOD-Net model, it preserves the color fidelity while also eliminating apparent smoke. On the other hand, our network and all of the compared methods still fail in heterogeneous smoke with high spatial variation in smoke density.

5. CONCLUSION

Surgery smoke removal is a critical issue for video guided surgery as mechanical solutions still require human labour. Therefore it is demanded from practitioners to have an automated digital solution where video frames are enhanced in real-time. That motivates us to approach the problem with neural networks. Although training step takes a long time, our network achieves 22 fps for color images of size 512x512 when tested on a single NVDIA Titan X GPU. To improve our network, we propose several ideas which will be realized in the future work:

1. Although we aim to improve perceptual image quality, MSE loss is used when updating the weights. It is aimed to utilize a perceptually relevant loss function during neural network training instead of MSE.
2. After $K(x)$ factor is estimated, smoke-free image is computed by pixel-wise linear operations applied to original image directly. Instead of directly using original image, we aim to utilize a denoising autoencoder to augment the method.

6. REFERENCES

[1] Camran Nezhat, Farr Nezhat, and Ceana Nezhat, *Nezhat's Video-Assisted and Robotic-Assisted Laparoscopy and Hysteroscopy*, Cambridge University Press, 2013.

[2] Congcong Wang, Rafael Palomar, and Faouzi Alaya Cheikh, "Stereo video analysis for instrument tracking in image-guided surgery," in *EUVIP*. IEEE, 2014, pp. 1–6.

[3] William L Barrett and Shawn M Garber, "Surgical smoke: a review of the literature," *Surgical Endoscopy*, vol. 17, no. 6, pp. 979–987, 2003.

[4] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng, "An all-in-one network for dehazing and beyond," *arXiv preprint arXiv:1707.06543*, 2017.

[5] Eric C Larson and Damon M Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.

[6] Kaiming He, Jian Sun, and Xiaoou Tang, "Single image haze removal using dark channel prior," *IEEE TPAMI*, vol. 33, no. 12, pp. 2341–2353, 2011.

[7] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE TIP*, vol. 25, no. 11, pp. 5187–5198, 2016.

[8] Adrian Galdran, Javier Vazquez-Corral, David Pardo, and Marcelo Bertalmío, "Enhanced variational image dehazing," *SIAM Journal on Imaging Sciences*, vol. 8, no. 3, pp. 1519–1546, 2015.

[9] Kevin Tchaka, Vijay M Pawara, and Danail Stoyanova, "Chromaticity based smoke removal in endoscopic images," in *Proc. of SPIE Vol*, 2017, vol. 10133, pp. 101331M–1.

[10] Alankar Kotwal, Riddhish Bhalodia, and Suyash P Awate, "Joint desmoking and denoising of laparoscopy images," in *ISBI*. IEEE, 2016, pp. 1050–1054.

[11] Ayush Baid, Alankar Kotwal, Riddhish Bhalodia, SN Merchant, and Suyash P Awate, "Joint desmoking, specularity removal, and denoising of laparoscopy images via graphical models and bayesian inference," in *ISBI*. IEEE, 2017, pp. 732–736.

[12] Xiongbiao Luo, A McLeod, Stephen Pautler, Christopher Schlachta, and Terry Peters, "Vision-based surgical field defogging," *IEEE TMI*, 2017.

[13] Earl J McCartney, "Optics of the atmosphere: scattering by molecules and particles," *New York, John Wiley and Sons, Inc., 421 p.*, 1976.

[14] Ko Nishino, Louis Kratz, and Stephen Lombardi, "Bayesian defogging," *IJCV*, vol. 98, no. 3, pp. 263–278, 2012.

[15] Codruta Orniana Ancuti and Cosmin Ancuti, "Single image dehazing by multi-scale fusion," *IEEE TIP*, vol. 22, no. 8, pp. 3271–3282, 2013.

[16] Hojatollah Yeganeh and Zhou Wang, "Objective quality assessment of tone-mapped images," *IEEE TIP*, vol. 22, no. 2, pp. 657–667, 2013.

[17] Khosro Bahrami and Alex C Kot, "A fast approach for no-reference image sharpness assessment based on maximum local variation," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 751–755, 2014.

[18] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," *ECCV*, pp. 746–760, 2012.

[19] Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 14–24, 2010.

[20] Danail Stoyanov, George P Mylonas, Fani Deligianni, Ara Darzi, and Guang Zhong Yang, "Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures," in *MICCAI*. Springer, 2005, pp. 139–146.

[21] Ken Perlin, "An image synthesizer," *ACM Siggraph Computer Graphics*, vol. 19, no. 3, pp. 287–296, 1985.

[22] Ken Perlin, "Improving noise," in *ACM TOG*. ACM, 2002, vol. 21, pp. 681–682.

[23] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang, "Single image dehazing via multi-scale convolutional neural networks," in *ECCV*. Springer, 2016, pp. 154–169.

[24] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[25] Marius Pedersen, "Evaluation of 60 full-reference image quality metrics on the cid: Iq," in *ICIP*. IEEE, 2015, pp. 1588–1592.

[26] S Giannarou, D Stoyanov, D Noonan, G Mylonas, J Clark, M Visentini-Scarzanella, P Mountney, and GZ Yang, "Hamlyn centre laparoscopic/endoscopic video datasets," 2012.