Towards a Video Quality Assessment based Framework for Enhancement of Laparoscopic Videos

Zohaib Amjad Khan[a], Azeddine Beghdadi[a], Faouzi Alaya Cheikh[b], Mounir Kaaniche[a], Egidijus Pelanis[c,d], Rafael Palomar[b], Åsmund Avdem Fretland[c,d,e], Bjørn Edwin[c,d,e], and Ole Jakob Elle[c,f]

[a]L2TI-Institut Galilée, Université Paris 13, Villetaneuse, France
[b]Norwegian Colour and Visual Computing Lab, NTNU, Gjøvik, Norway
[c]The Intervention Centre, Oslo University Hospital – Rikshospitalet, Oslo, Norway
[d]Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[e]Department of HPB Surgery, Oslo University Hospital – Rikshospitalet, Oslo, Norway
[f]Department of Informatics, University of Oslo, Oslo, Norway

ABSTRACT

Laparoscopic videos can be affected by different distortions which may impact the performance of surgery and introduce surgical errors. In this work, we propose a framework for automatically detecting and identifying such distortions and their severity using video quality assessment. There are three major contributions presented in this work (i) a proposal for a novel video enhancement framework for laparoscopic surgery; (ii) a publicly available database for quality assessment of laparoscopic videos evaluated by expert as well as non-expert observers and (iii) objective video quality assessment of laparoscopic videos including their correlations with expert and non-expert scores.

Keywords: subjective evaluation, laparoscopic video, video quality assessment, distortion classification

1. INTRODUCTION

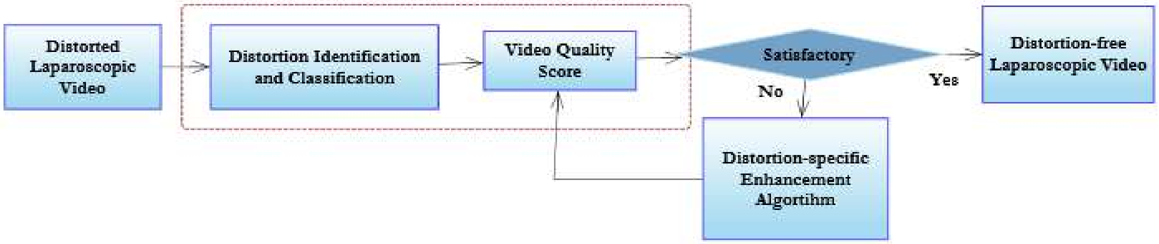
A satisfactory video quality is an important requirement for achieving optimal conditions for laparoscopic surgery. The distortions in a laparoscopic video not only affect a surgeon's visibility but also degrade the results of subsequent computational tasks in robot-assisted surgery and image-guided navigation systems.[1] Examples of such tasks are segmentation,[2,3] instrument tracking,[4,5] and augmented reality.[6]

Laparoscopic videos may be affected by different kinds of distortions during the surgery, resulting in a loss of visual quality. These are mainly a result of technical problems in the equipment[7] or side-effects of the instruments being used (e.g. smoke with diathermy). To deal with such problems, most of the suggested prevalent solutions rely on making some changes to the technical equipment using one of the many available troubleshooting options as also highlighted in Ref. 8. However, all such solutions are time-consuming and may not always solve the problem at hand requiring eventually a specialist technician or a change in apparatus.[7]

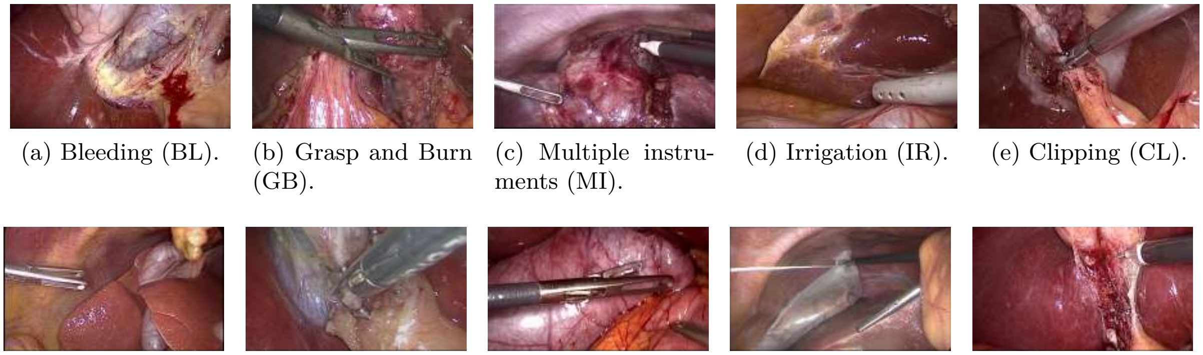
In this work, we propose a computational framework for automatic detection and correction of video quality for laparoscopic surgery (Figure 1). The proposed framework consists of a video quality assessment (VQA) module followed by an enhancement module. This work solely focuses on the video quality assessment part (dotted red area in Figure 1) which is composed of two stages namely distortion detection/classification and video quality score evaluation. Such hybrid two-step quality assessment techniques are not new and have already been proposed for natural images.[9]

Quality assessment of videos, if done subjectively, is time-consuming and hence not feasible. In order to assess video quality automatically, objective metrics are needed. However, the effectiveness of a designed objective

Further author information: (Send correspondence to Z.A. Khan)
Z.A. Khan: E-mail: zohaibamjad.khan@univ-paris13.fr,
A. Beghdadi: E-mail: beghdadi@univ-paris13.fr


Distorted Laparoscopic Video
Distortion Identification and Classification
Video Quality Score
Satisfactory
No
Yes
Distortion-free Laparoscopic Video
Distortion-specific Enhancement Algortihm

Figur 1: Flowchart of the proposed framework for laparoscopic video enhancement with quality control



(a) Bleeding (BL). (b) Grasp and Burn (GB). (c) Multiple instruments (MI). (d) Irrigation (IR). (e) Clipping (CL).

(f) Stretching away (SA). (g) Cutting (CU). (h) Grasping and stretching forward (SF). (i) Organ Extraction (OE). (j) Burning (BU).

Figur 2: One frame from each of the reference videos in the LVQ database.

metric can only be evaluated by using a database of videos annotated with subjective scores.[10] Unfortunately, to the best of our knowledge, there is no such database of laparoscopic videos available publicly. Hence, there is currently a big gap to be filled in the field of medical visual quality assessment. This work also aims to fill this gap by proposing a new database which is dedicated to laparoscopic video quality assessment (Available at: LVQ Database) [*]

2. LAPAROSCOPIC VIDEO QUALITY (LVQ) DATABASE

Our database called the Laparoscopic Video Quality (LVQ) database consists of a total of 10 reference videos, each of 10 seconds. Each reference video is distorted by five different kinds of distortions with four different levels, resulting in a total of 200 videos. The resolution of the videos is 512 × 288 with a 16:9 aspect ratio and a frame-rate of 25 fps. Moreover, we have used uncompressed avi format for the videos so as to avoid any kind of unwanted compression artefacts like blocking. In the following sections, we describe in more details the construction of our database.

2.1 Selection of Videos

For the database, ten different videos of laparoscopic cholecystectomy are selected as reference. These videos are extracted from Cholec80 dataset[11] and are shown in Figure 2. The selection of the videos is made with an attempt to include maximum possible variations of scene content and temporal information. For scene content, ten different categories are chosen as illustrated in Figure 2. These are bleeding (BL), grasping and burning (GB), multiple instruments (MI), irrigation (IR), clipping (CL), stretching away (SA), cutting (CU), stretching forward (SF), organ extraction (OE) and burning (BU).

[*] URL: https://drive.google.com/file/d/1SoONeacp9vvihTY7zmWssG_cnVzx16oq/view?usp=sharing.

2.2 Creation of Distorted Videos

We have chosen five different distortions for our database. These distortions, which are among the most common affecting a laparoscopic video, are the smoke, noise, uneven illumination, blur due to defocus and blur due to motion. In order to simulate each of these distortions, we have applied appropriate mathematical models to every frame of the reference video. For generating four levels of severity for each distortion, we have modified the relevant parameters of these models. For this initial work, we have not considered time-variatons in distortions. Hence, in each video there is one single distortion at same level throughout.

We have used MATLAB to simulate all the distortions with different levels for our database. For defocus blur, a symmetric low-pass Gaussian filter was applied to each video frame. The filter size and the standard deviation were changed to generate different levels of this distortion. For motion blur, MATLAB motion filter was used with variations in filter length parameter to generate different severity levels. Similarly, built-in MATLAB function for noise was used to add additive white Gaussian noise. In this case, variance of the Gaussian model was adjusted to generate multiple levels of noisy videos.

For uneven illumination, a special grayscale mask was generated having a circular bright region of high intensity values. The areas surrounding the circular region were generated with decreasing intensities which were attenuated as a function of distance from the center of the bright region. The multiplication of this mask with the original frame gave us the unevenly illuminated distorted frame. By changing the two parameters of the center location of the bright region and its area, we generated four different levels for unveven illumination.

In order to generate smoke, we have used a well-known method of video editing called the screen blending. In this technique, a smoke-only video having a black background is combined with the reference video in such a way that black areas produce no change to the original video while the brighter areas overlay the original ones. Using different opacity levels for the smoke video we have generated four different levels of severity for smoke.

2.3 Subjective Tests

For the subjective testing, we have used pairwise-comparison protocol.[12,13] For each observer, we randomly displayed all possible pair combinations of distorted videos, such that each pair had two videos from the same category and the same distortion type, with only difference being the severity level. This corresponded to 6 pair-wise comparisons per reference video for each distortion type.

The observers had the choice to give an equal score to the videos if they perceived so. For each comparison, the preferred video was given one point. In case of an equal choice, a score of 0.5 was given to each displayed video. The observers were shown each video once and they had the choice to see the video again if required. In the end, the scores for each video from all the observers were added. Finally the Mean Opinion Score (MOS) for the i-th video was obtained by averaging the total score for that video over number of observers N.[14]

$$MOS_i = \frac{1}{N}\sum_{j=1}^{N} score_{ij} \tag{1}$$

In order to perform the subjective tests, a calibrated 24.1 inch LCD monitor was used. The observers were forced to perform the experiments at a fixed distance of twice the screen height which is equivalent in our case to be 4.5 times the image height of the image. The setup for the subjective tests is shown in Figure 3. All the observers had either normal vision or corrected to normal vision and they were undergone a pre-screening procedure for color vision and visual acuity.

3. LAPAROSCOPIC VIDEO QUALITY ASSESSMENT

In order to apply the pertinent enhancement method with suitable parameters in our proposed video enhancement framework, VQA module should be able to detect the type of distortion affecting a video as well as its severity. For this reason, our VQA consists of a distortion identification step followed by the quality score estimation.

Figur 3: Setup for subjective tests

3.1 Distortion Classification

As a first step, for each kind of distortion, we have chosen a distortion-specific classification method. In our selection of these methods, we have gone for no-reference, opinion-unaware, accurate and computationally less expensive methods to allow for real-time performance. Below are the details of the classification methods used for each kind of distortion.

3.1.1 Motion and defocus blur

For the two kinds of blur, we have used Perceptual Blur Index (PBI)[15] with threshold as the classifier. PBI is a quality metric for estimating the level of blur in an image. It is based on the way Human Visual System (HVS) perceives addition of blur to an already blurred image and to a sharp one differently . The perceptual difference is more pronounced for the latter case. It is defined in terms of the difference between total radial energy of the input image $RE(w)$ and that of its binomial filtered version $RE_f(w)$ as

$$PBI = \log(\frac{1}{w_{max}} \sum_{w} |RE(w) - RE_f(w)|) \tag{2}$$

where w_{max} is the maximal frequency.

3.1.2 Smoke

In order to detect if there is smoke in a video, we have used Saturation Analysis (SAN) classifier.[16] SAN classifier uses the histogram of saturation channel of a frame to detect smoke. If the majority of bin values in histogram $hist$ are below the chosen threshold t_c, the video frame is classified to have smoke in it. The threshold used is $t_c = 0.35$ as suggested in the original work.[16] The probability of an image having smoke $p_{(S)}$ and no smoke $p_{(NS)}$ are therefore defined as

$$p_{(S)} = \frac{1}{|hist|} \sum_{\substack{i=0 \\ b \in hist \\ t \leq t_c}} b_i \tag{3}$$

$$p_{(NS)} = 1 - p_{(S)} \tag{4}$$

where b_i is the i-th bin value of the histogram $hist$.

3.1.3 Noise

For noise classification we have chosen the fast noise variance estimator[17] with threshold. In this method, the standard deviation of additive white Gaussian noise in an image is estimated using a noise estimation mask. This suggested mask, M_N has been generated using a difference of two 3×3 masks, each approximating the Laplacian of an image. For an image I with width W and height H, the estimated standard deviation σ_n of noise is estimated as:[17]

$$\sigma_n = \sqrt{\frac{\pi}{2}} \frac{1}{6(W-2)(H-2)} \sum_{x,y} |I(x,y) * M_N| \tag{5}$$

3.1.4 Uneven illumination

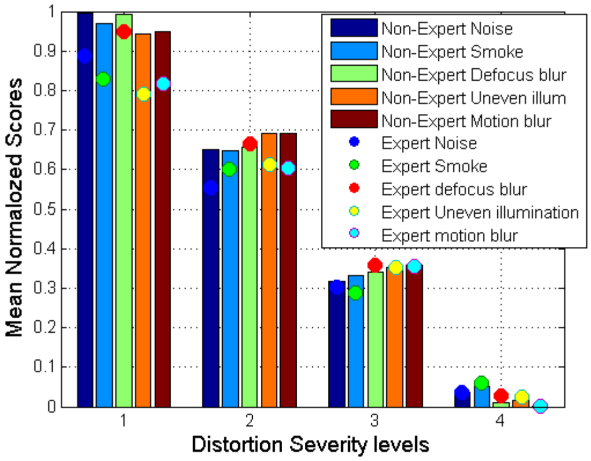
In order to detect whether a video is affected by uneven illumination or not, we have developed a novel classifier which makes use of statistics of the luminance component of an image. For an unevenly illuminated laparoscopic video frame, there are some dark regions in the image which tend to increase the range of values for the luminance component in an image, while reducing the mean luminance value of the image at the same time. Making use of these trends, we have proposed a new classifier that uses a threshold on the Luminance Mean to Range (LMR) ratio, defined simply as the ratio of mean luminance value to that of the range of luminance values in an image. For an image with N_p pixels and with luminance component Y, this index can be defined as

$$LMR = \frac{\frac{1}{N_p}\sum_{i=0}^{N_p} Y_i}{max(Y) - min(Y)} \tag{6}$$

An image with a LMR value smaller than a pre-defined threshold can be classified to have been affected by uneven illumination.

3.2 Video Quality Score

Once the distortion is identified in a video, we can evaluate its severity using a quality score. To this effect, we have selected 3 different metrics which are often used as benchmarks for natural images and videos. These are PSNR, SSIM[18] and VIF.[19] However, for laparoscopic videos, there is usually no ground truth available and a no-reference (NR) metric makes more sense. For this reason, we have also included NR metrics. However, due to the limited number of good NR metrics for videos, we have only selected one of the more recent NR metrics dedicated to opinion-unaware VQA called VIIDEO.[20] We have also included two NR image quality metrics BRISQUE[21] and NIQE.[22] For both of these, we have used the mean metric value from all frames as the score for the video.



Figur 4: Comparison of subjective scores for experts and non-experts

Tabell 1: PLCC for **non-expert** scores in LVQ Database (best two values in bold for each column)

Metric	**Noise**	**Defocus Blur**	**Motion Blur**	**Uneven illumination**	**Smoke**	**Overall**
PSNR	**0.9968**	0.8166	0.8199	0.9561	**0.9811**	0.6054
SSIM	0.9690	0.7388	**0.8861**	**0.9926**	0.9165	**0.6123**
VIF	**0.9925**	**0.9764**	**0.9713**	**0.9919**	**0.9853**	**0.6267**
BRISQUE	0.9803	0.9646	0.4090	0.3142	0.3735	0.4593
NIQE	0.9783	**0.9880**	0.7704	0.6618	0.3238	0.4242
VIIDEO	0.8749	0.3549	0.4998	0.3983	0.4214	0.3842

Tabell 2: SROCC for **non-expert** scores in LVQ Database (best two values in bold for each column)

Metric	**Noise**	**Defocus Blur**	**Motion Blur**	**Uneven illumination**	**Smoke**	**Overall**
PSNR	0.9594	0.7773	0.8163	0.9372	**0.9439**	0.5775
SSIM	0.9509	0.7157	**0.8941**	**0.9502**	0.8987	**0.5914**
VIF	**0.9636**	**0.9417**	**0.9433**	**0.9391**	**0.9316**	**0.6228**
BRISQUE	0.9571	0.9332	0.3564	0.2980	0.4041	0.4304
NIQE	**0.9640**	**0.9514**	0.6101	0.5416	0.3589	0.3731
VIIDEO	0.8600	0.3138	0.379	0.3888	0.3866	0.3416

4. RESULTS AND DISCUSSION

In total, thirty non-expert and ten expert observers performed the subjective tests for the database. Both the expert and non-expert observers were considered as two separate groups. For each group, outliers were first detected based on non-transitivity. This corresponded to one subject in each group. The preference matrices for remaining subjects in the two groups were then compiled and aggregated to obtain subjective scores.

Figure 4 shows a comparison between expert and non-expert mean normalized scores for LVQ database. From the figure, we can clearly see how experts perceive quality much differently for all distortions except for defocus blur. The difference is more pronounced for less distorted videos (levels 1 and 2) suggesting how even the slightest level of distortion affects the perception of a video for experts (who are more task-oriented).

To evaluate the performance of our selected classification methods, all the videos in our database were passed through these classifiers. The results obtained for classification accuracies were: smoke - 87%; motion blur - 89%;

Tabell 3: PLCC for **expert** scores in LVQ Database (best two values in bold for each column)

Metric	**Noise**	**Defocus Blur**	**Motion Blur**	**Uneven illumination**	**Smoke**	**Overall**
PSNR	**0.9939**	0.8146	0.8226	0.9452	**0.9777**	**0.6853**
SSIM	0.9706	0.7358	**0.8827**	**0.9847**	0.9116	0.5732
VIF	**0.9896**	**0.9806**	**0.9708**	**0.9878**	**0.9808**	**0.5909**
BRISQUE	0.9761	0.9623	0.4208	0.2973	0.4009	0.4434
NIQE	0.9741	**0.9883**	0.7836	0.6655	0.4301	0.4407
VIIDEO	0.8658	0.3498	0.5136	0.4035	0.4195	0.3744

Tabell 4: SROCC for **expert** scores in LVQ Database (best two values in bold for each column)

Metric	Noise	Defocus Blur	Motion Blur	Uneven illumination	Smoke	Overall
PSNR	0.9579	0.7836	0.7977	0.9530	**0.9478**	**0.6914**
SSIM	0.9435	0.7320	**0.8802**	**0.9580**	0.8817	**0.5653**
VIF	**0.9592**	**0.9555**	**0.9376**	**0.9534**	**0.9459**	0.5642
BRISQUE	0.9527	0.9355	0.3994	0.2634	0.4355	0.3842
NIQE	**0.9594**	**0.9443**	0.7028	0.5605	0.3382	0.3674
VIIDEO	0.8822	0.3023	0.3915	0.4281	0.4416	0.3334

defocus blur - 91.5%; noise - 100% and uneven illumination classifiers - 88.5%.

Furthermore, in order to assess whether an existing objective video quality metric correlates well or not with subjective scores, Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) were evaluated for the metric scores after performing a non-linear regression with a 5-parameter logistic function. From Tables 1 and 3, we can see that none of the objective metrics perform well when overall correlations are evaluated, with maximum PLCC value of 0.6267 with VIF for non-experts and a value of 0.6853 with PSNR for experts.

However, with individual distortion types, VIF correlates much better with subjective scores as compared to others for both groups and for all the distortions. Among the NR metrics, both NIQE and BRISQUE give good results for noise and defocus blur, with NIQE being better of the two for motion blur and uneven illumination. However, VQA specific method VIIDEO performs poorly for all distortions except for the noise.

All these results are very significant as they imply that none of these metrics are generic or non-distortion specific for the kind of videos and distortions encountered in medical domain. Moreover, these results also show a difference with respect to their correlation with expert and non-expert scores. To be more specific, if we compare the results of experts and non-experts, we can see that generally all the metrics tend to correlate better with non-expert opinion as compared to expert opinion.

5. CONCLUSION

In this work, we have proposed a novel computational framework for laparoscopic video enhancement based on video quality assessment. Especially, we have taken a major initiative for quality assessment of laparoscopic videos by creating a database with subjective quality scores not only from normal observers but also from medical experts. Our initial results show that the existing NR metrics for video quality assessment are not sufficient especially in context of laparoscopic videos. Moreover, we have observed that experts and non-experts differ in their opinions on video quality assessment and new no-reference metrics are required to model expert opinion. In this regards, the constructed LVQ database is an important step to address the above challenges in a future work and in facilitating development of new VQA metrics in the medical imaging context.

ACKNOWLEDGMENTS

This research work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 722068. We would also like to thank all the observers who took part in the subjective experiments especially all the medical experts from Oslo University Hospital (Rikshospitalet), Norway.

REFERENCES

[1] Sanchez-Gonzalez, P., Cano, A. M., Oropesa, I., Sanchez-Margallo, F. M., Pozo, F. D., Lamata, P., and Gómez, E. J., "Laparoscopic video analysis for training and image-guided surgery," *Minimally Invasive Therapy & Allied Technologies* **20**(6), 311–320 (2011).

[2] Bodenstedt, S., Ohnemus, A., Katic, D., Wekerle, A.-L., Wagner, M., Kenngott, H., Müller-Stich, B., Dillmann, R., and Speidel, S., "Real-time image-based instrument classification for laparoscopic surgery," *arXiv preprint arXiv:1808.00178* (2018).

[3] Voros, S., Long, J.-A., and Cinquin, P., "Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders," *The International Journal of Robotics Research* **26**(11-12), 1173–1190 (2007).

[4] Bouget, D., Allan, M., Stoyanov, D., and Jannin, P., "Vision-based and marker-less surgical tool detection and tracking: a review of the literature," *Medical image analysis* **35**, 633–654 (2017).

[5] Zhou, J. and Payandeh, S., "Visual tracking of laparoscopic instruments," *Journal of Automation and Control Engineering Vol* **2**(3) (2014).

[6] Bernhardt, S., Nicolau, S. A., Soler, L., and Doignon, C., "The status of augmented reality in laparoscopic surgery as of 2016," *Medical image analysis* **37**, 66–90 (2017).

[7] Verdaasdonk, E. G., Stassen, L. P., van der Elst, M., Karsten, T. M., and Dankelman, J., "Problems with technical equipment during laparoscopic surgery," *Surgical endoscopy* **21**(2), 275–279 (2007).

[8] Siddaiah-Subramanya, M., Nyandowe, M., and Tiang, K. W., "Technical problems during laparoscopy: a systematic method of troubleshooting for surgeons," *Innovative Surgical Sciences* **2**(4), 233–237 (2017).

[9] Chetouani, A., Beghdadi, A., and Deriche, M., "A hybrid system for distortion classification and image quality evaluation," *Signal Processing: Image Communication* **27**(9), 948–960 (2012).

[10] Winkler, S., "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing* **6**(6), 616–625 (2012).

[11] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., and Padoy, N., "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging* **36**(1), 86–97 (2017).

[12] ITU-T, R., "P910," *Subjective video quality assessment methods for multimedia applications* (2008).

[13] Qureshi, M. A., Beghdadi, A., and Deriche, M., "Towards the design of a consistent image contrast enhancement evaluation measure," *Signal Processing: Image Communication* **58**, 212–227 (2017).

[14] Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication* **30**, 57–77 (2015).

[15] Chetouani, A., Beghdadi, A., and Deriche, M., "A new reference-free image quality index for blur estimation in the frequency domain," in [*2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*], 155–159, IEEE (2009).

[16] Leibetseder, A., Primus, M. J., Petscharnig, S., and Schoeffmann, K., "Real-time image-based smoke detection in endoscopic videos," in [*Proceedings of the on Thematic Workshops of ACM Multimedia*], 296–304 (2017).

[17] Immerkaer, J., "Fast noise variance estimation," *Computer vision and image understanding* **64**(2), 300–302 (1996).

[18] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing* **13**(4), 600–612 (2004).

[19] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE Transactions on image processing* **15**(2), 430–444 (2006).

[20] Mittal, A., Saad, M. A., and Bovik, A. C., "A completely blind video integrity oracle," *IEEE Transactions on Image Processing* **25**(1), 289–300 (2015).

[21] Mittal, A., Moorthy, A. K., and Bovik, A. C., "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing* **21**(12), 4695–4708 (2012).

[22] Mittal, A., Soundararajan, R., and Bovik, A. C., "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters* **20**(3), 209–212 (2012).