Sria Biswas* and Rohini Palanisamy

Enhancing No Reference Laparoscopic Video Quality Assessment with Evolutionary ANFIS

https://doi.org/10.1515/cdbme-2024-2021

Abstract: Distortions in laparoscopic videos affect surgeon visibility and surgical precision, underscoring the need for sustained high video quality. This study presents a real-time laparoscopic video quality assessment algorithm independent of reference content availability. Statistical parameters derived from luminance, local binary pattern and motion-vector maps of video frames are observed to effectively discern distortion types and severities. These parameters are used to train an evolutionary adaptive neuro-fuzzy inference system (ANFIS) end-to-end with subjective score labels. Training and validation loss curves saturate at the 85th epoch, demonstrating the model's efficient data fitting capability. Performance comparison with other state-of-the-art methods reveals superior results, with high correlation scores of 0.9989 and 0.9446 for experts and 0.9956 and 0.9847 for non-experts, alongside low root mean square errors of 0.0828 and 0.1685 for expert and non-experts, respectively. The model accurately replicates the expert and non-expert perceptual opinions, encouraging future research in stereoscopic, augmented, and virtual reality data.

Keywords: Laparoscopic Surgery, Video Quality Assessment, Evolutionary ANFIS, No Reference, Feature Selection.

1 Introduction

Recent years have witnessed a surge in the popularity of laparoscopic surgeries, driven by their various advantages, such as reduced trauma, faster recovery times, and enhanced patient outcomes. Biomedical engineering advancements offer promising avenues for integrating cutting-edge technologies like robotics, stereoscopic 3D (S3D), virtual reality (VR), and augmented reality (AR) into laparoscopic procedures, promising safer, more accessible, and cost-effective options for patients worldwide. However, sustaining this growth demands the maintenance of high-quality videos during laparoscopy.

These streams serve as indispensable guides for surgeons during minimally invasive procedures, facilitating precise interventions inside surgical sites, identification of disease biomarkers, real-time feedback provision, and post-operative analysis. Yet, these videos are susceptible to degradation from factors like surgical smoke, uneven illumination, and lens defects. Therefore, the development of no-reference (NR) laparoscopic video quality assessment (LVQA) algorithms becomes imperative, enabling real-time quality estimation without relying on undistorted reference content, which is a formidable challenge in the medical domain.

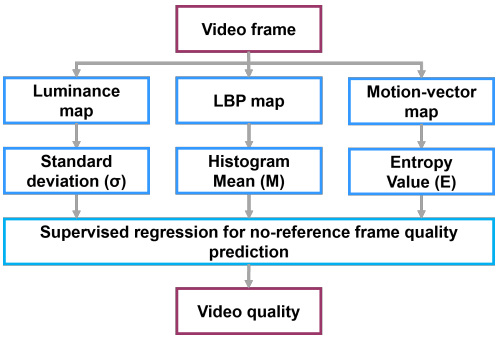

Fig. 1: Flowchart of the proposed NR LVQA algorithm

Previous works have employed conventional objective metrics, image quality distance measures, joint feature dependencies, or stacked framework deep features to compute the quality of video frames [1–4]. However, a unified feature combination derived from locally extracted spatial, texture, and temporal features has not been explored. Since human perception assesses scene quality based on multiple relevant features, it is imperative to consider a combination to replicate this perception accurately. Additionally, statistically extracted handcrafted features offer the advantage of customization to enhance frame quality estimation. This is particularly crucial for laparoscopic videos, which typically exhibit limited temporal activity, small visual regions, and heightened textural activity from tissues, among other factors.

Once features are selected, regression frameworks are employed to map these features to their perceptual quality labels. Support vector regressors, adaptive neuro-fuzzy inference systems (ANFIS), and residual networks have all been utilized previously [2–4]. Among these, ANFIS integrates artificial neural network (ANN) with the Takagi–Sugeno fuzzy inference system (FIS) by combining fuzzy logic and neural networks concepts. This enables it to effectively model complex, non-linear relationships. Additionally, its adaptive learning capability allows it to adjust parameters based on inputs, enhancing regression performance over time.

In this work, an NR LVQA algorithm is developed by leveraging the end-to-end (E2E) training of an evolutionary

***Corresponding author: Sria Biswas, Rohini Palanisamy,** Department of Electronics and Communications Engineering, Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, Chennai, Tamil Nadu, Pin code - 600127, India, e-mail: ec21d0002@iiitdm.ac.in, rohinip@iiitdm.ac.in

ANFIS framework with locally extracted spatial, texture, and temporal feature components at frame-level to estimate video quality. The network directly outputs frame quality, and the video quality is obtained by taking the mean of all predictions.

2 Methodology

Figure 1 illustrates the flowchart of the proposed algorithm and the stages are discussed below.

2.1 Database Selection

The openly accessible Laparoscopic Video Quality (LVQ) database comprises of 10 reference and 200 distorted videos designed by simulating 4 distortion types, namely motion blur (MB), defocus blur (DB), additive white Gaussian noise (WN), smoke (SM) and uneven illumination (UI), at distortion levels (1, 2, 3, 4) [1]. The videos have 512×288 resolution, 25 fps frame rate and 10 seconds duration. The subjective quality mean opinion scores (MOS) of each video is available for 10 experts and 30 non-experts.

2.2 Feature Extraction

Human vision perceives scene quality through retinal processing of perceptual components like brightness, contrast, structure, edges, and movement dynamics. To replicate this perception, spatial, texture, and temporal feature extraction is conducted at the frame-level as follows:

1. **Spatial Feature:** Luminance maps are generated by converting RGB frames into grayscale images and removing their hue and saturation values. The standard deviation (σ) of this map is calculated using the formula: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$, where N, x_i and μ represent the total number of pixels, intensity of each pixel, and mean intensity value, respectively.
2. **Texture Feature:** Local binary pattern (LBP) descriptor is obtained by comparing the central pixel intensity to its neighbors, assigning binary values based on a threshold, and summing these binary values. The mean value of the LBP histogram ($\boldsymbol{M}$) is computed as: $\boldsymbol{M} = \frac{1}{N}\sum_{i=1}^{n}(f_i \times b_i)$, where N is total number of pixels, and f_i and b_i indicate frequency of occurrences and value of i_{th} bin.
3. **Temporal Feature:** Motion-vector maps are generated using the Lucas-Kanade method to estimate optical flow between succesive frames. The magnitude of the motion vectors is calculated, followed by estimating their entropy values ($\boldsymbol{E}$) given by, $\boldsymbol{E} = -\sum_{i=1}^{m} p_i \log_2(p_i)$, where m is number of bins used for discretizing the magnitude values and p_i denotes probability of occurrence of bin i.

2.3 ANFIS Architecture

Figure 2 shows the ANFIS architecture comprising of 5 layers which are outlined as follows:

1. **Fuzzification Layer:** In Layer-1, crisp inputs are transformed into fuzzy sets using rules for representing the degree of membership of inputs to each fuzzy set.
2. **Rule Layer:** In Layer-2, the firing strength (w_i) of a rule is computed by the product of the fuzzified inputs.
3. **Normalization Layer:** In Layer-3, the normalized firing strengths ($\bar{w}_i$) are estimated by the ratio of firing strength of i^{th} rule to sum of firing strengths of all rules.
4. **Consequent Layer:** In Layer-4, $\bar{w}_i$ is multiplied to constant values associated with the fuzzy rules (f_i) to generate a linear function of the inputs.
5. **Aggregation Layer:** In Layer-5, the final output is calculated as sum of all values obtained from previous layer.

2.4 Frame Quality Prediction

The ANFIS conducts supervised non-linear regression to predict frame quality scores, utilizing both video-level expert and non-expert MOS training labels individually [3]. The dataset is split into non-overlapping train-validation-test in a 60:20:20 ratio, and the (σ,$\boldsymbol{M}$,$\boldsymbol{E}$) feature vector is used as input, while omitting reference features to enable NR analysis. The Gaussian membership function is utilized for the fuzzification process, alongside fuzzy c-means clustering and particle swarm optimization (PSO) during E2E training for 85 epochs with an initial step size of 0.009. The computational time involved from start till frame quality prediction is approx. 0.48 seconds.

2.5 Video Quality Estimation

The predicted frame quality scores, being time-varying in nature, are averaged across all frames to derive the overall video quality score, thereby delivering predicted expert and non-expert MOS.

3 Result and Discussions

The representative reference and highest distorted frames for all distortion types from the LVQ database are depicted in Figure 3 (row 1). This dataset encompasses a diverse range of distortions and distortion levels commonly encountered by surgeons during laparoscopy in real-life scenarios.

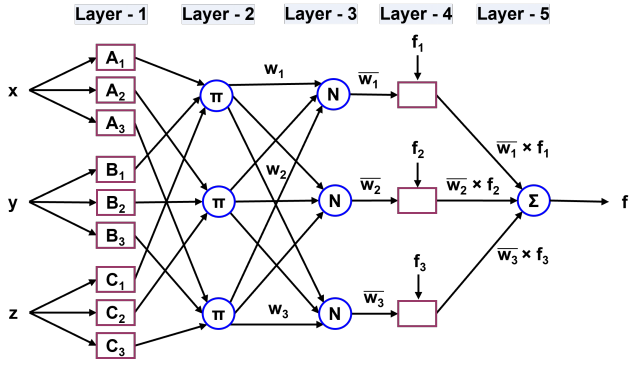

Fig. 2: 3 inputs and 1 output Takagi-Sugeno ANFIS architecture

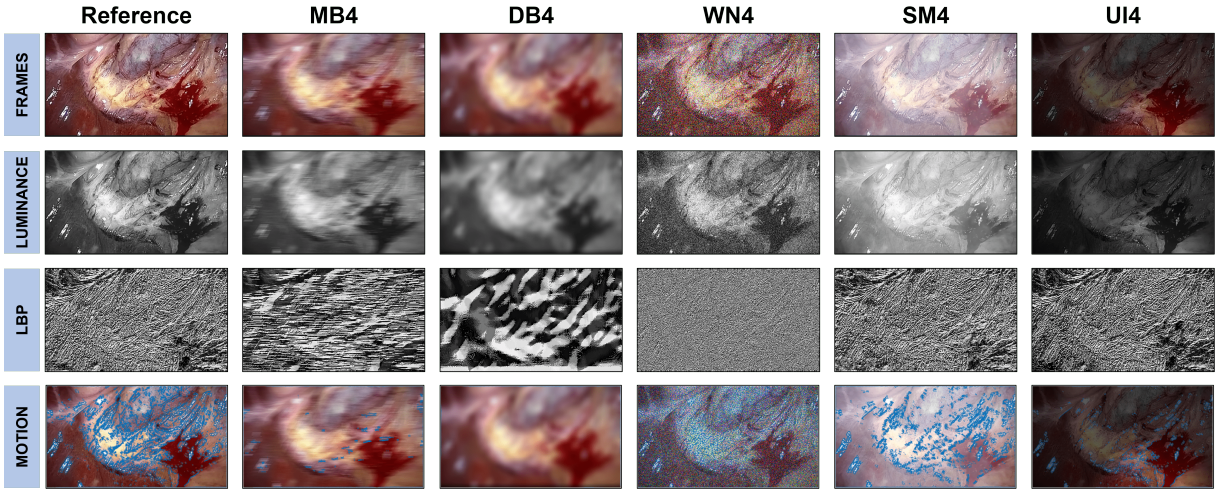

Fig. 3: Representative frames from reference and corresponding distorted videos at the highest distortion level

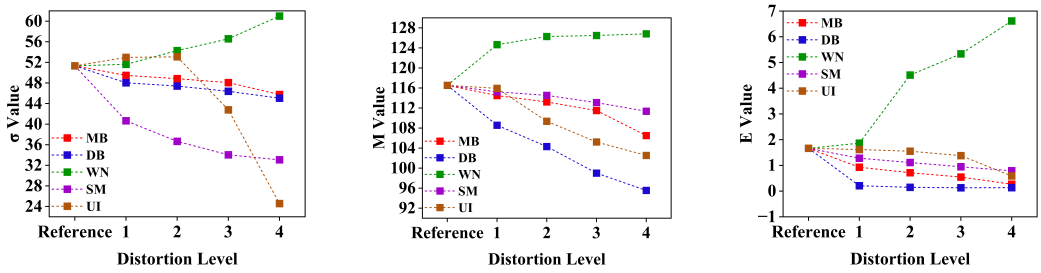

Fig. 4: Visualization of (σ, M, E) feature distribution for reference and four distortion levels across five distortion types

Luminance, LBP and motion-vector maps for the corresponding frames are represented in rows 2, 3 and 4 of Figures 3. The luminance maps of distorted frames display regions of over-exposure or shadows, resulting in artificially heightened contrast levels, diminished sharpness, and clarity. This can obscure biomarkers and finer tissue structures, leading to judgemental errors by a surgeon. Meanwhile, LBP maps exhibit a general loss of texture and pattern details crucial for identifying edges and structural information, which are vital for the surgeon to perceive scene depth and distinguish between tissue structures. Motion-vector maps indicate increased temporal activity for noise distortions, while other distortions display reduced temporal activity, ultimately leading to motion loss. Overall, these chosen features effectively apprehend the information loss caused by different distortions, demonstrating their relevance in perceptual assessment.

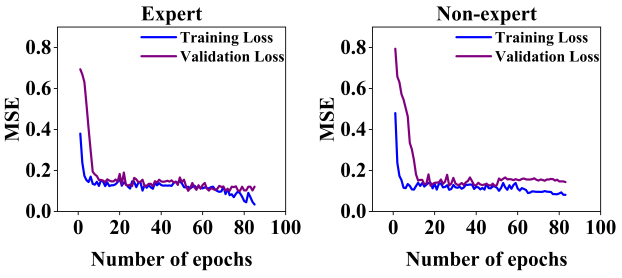

Fig. 5: Visualisation of training and validation loss curves

The generated feature maps undergo various statistical estimations to quantify their contribution in capturing underlying quality defining characteristics. The standard deviation (σ) of the grayscale image signifies scene contrast by measuring the deviation of pixel intensities from the mean intensity value. Additionally, the mean value (M) of the LBP histogram gauges the average distribution of local texture patterns present in the frame to record texture diversity. Motion entropy (E) is estimated to represent dynamic shifts and subtle motion patterns by accounting for pixel-level motion, thus facilitating the detection of even minor movements. The distribution of these features is illustrated in Figure 4. It is evident that these statistical parameters can effectively differentiate between various distortion types and levels, demonstrating their potential as distortion discriminative inputs for a regression network.

The ANFIS framework predicts both expert and non-expert frame quality based on the two training labels. Regression analysis is conducted by estimating the mean square error (MSE) to determine how well the model fits the provided data. Figure 5 depicts the training and validation loss curves for both cases, showing a gradual decrease in loss with each epoch, eventually plateauing around the 85th iteration at 0.0344 and 0.1210 for experts, and 0.0811 and 0.1447 for non-experts.

The proposed algorithm performance is compared to other state-of-the-art NR image QA and VQA methods using Pear-

Tab. 1: Performance comparison of proposed algorithm in terms of PLCC and SROCC scores for **expert** training labels

Type	Algorithm	PLCC					SROCC				
		MB	DB	WN	SM	UI	MB	DB	WN	SM	UI
NR IQA	BRISQUE [5]	0.7419	0.9244	0.9743	0.2369	0.4165	0.8788	0.7389	0.9464	0.2481	0.4255
	NIQE [6]	0.3137	0.7903	0.9762	0.0112	0.2960	0.2376	0.8046	0.9560	0.0123	0.1856
	ILNIQE [7]	0.3915	0.8181	0.9793	0.2195	0.4539	0.3010	0.7965	0.9495	0.2512	0.3831
NR VQA	VIIDEO [8]	0.3498	0.5136	0.8658	0.4195	0.4035	0.3915	0.3023	0.8822	0.4416	0.4281
	TLVQM [9]	0.7493	0.7066	0.7651	0.6321	0.5986	0.7031	0.7386	0.7591	0.6326	0.5998
	Proposed	0.9858	0.9937	0.9950	0.9989	0.9851	0.9446	0.9416	0.8863	0.8924	0.9019

Tab. 2: Performance comparison of proposed algorithm in terms of PLCC and SROCC scores for **non-expert** training labels

Type	Algorithm	PLCC					SROCC				
		MB	DB	WN	SM	UI	MB	DB	WN	SM	UI
NR IQA	BRISQUE [5]	0.4090	0.9646	0.9803	0.3735	0.3142	0.3564	0.9332	0.9571	0.4041	0.2980
	NIQE [6]	0.7704	0.9880	0.9783	0.3238	0.6618	0.6101	0.9514	0.9640	0.3589	0.5416
NR VQA	VIIDEO [8]	0.4998	0.3549	0.8749	0.4214	0.3983	0.3790	0.3138	0.8600	0.3866	0.3888
	Proposed	0.9856	0.9831	0.9956	0.9779	0.9842	0.9847	0.9510	0.8896	0.9478	0.9208

Tab. 3: Proposed algorithm performance using RMSE scores

Training Label	MB	DB	WN	SM	UI
Expert	0.3008	0.2007	0.1800	0.0828	0.3044
Non-expert	0.2087	0.1685	0.3027	0.3700	0.3165

son's Linear Correlation Coefficient (PLCC) and Spearman's Rank Order Correlation Coefficient (SROCC). These metrics enable straightforward comparison and validation of predicted objective results, as demonstrated in tables 1 and 2. The proposed algorithm achieves high correlation scores for both categories, outperforming other NR methods and indicating strong coherence with subjective MOS values. These NR methods were specifically developed for natural videos, primarily relying on spatial components to measure quality scores. Meanwhile, the proposed model also incorporates texture and temporal cues from laparoscopic scenes, enhancing quality assessment and improving performance. The proposed algorithm results are also examined for different distortion types using root mean square error (RMSE) values, as shown in table 3. It demonstrates overall low RMSE values, signifying a reduced error margin between predicted and subjective MOS.

4 Conclusion

This paper introduces a NR LVQA algorithm employing an evolutionary ANFIS regressor and statistical analysis of frame-level feature maps. The standard deviation, histogram mean, and entropy values of luminance, LBP, and motion-vector maps respectively, is seen to effectively differentiate between various types and intensities of distortions, making them suitable inputs for ANFIS training. The MSE loss curves for both training and validation data decrease consistently with increasing epochs until the 85^{th} iteration, indicating the model effectiveness in predicting data. High PLCC and SROCC scores, coupled with low RMSE values, demonstrate the efficiency of the proposed model in delivering robust and competitive performance compared to other NR IQA and VQA methods. These results have the potential to foster further research in integrating the presented findings into S3D, AR, and VR simulations for real-time laparoscopy and surgical training applications.

Author Statement: **Research funding:** Authors state no funding involved. **Conflict of interest:** Authors state no conflict of interest. **Informed consent:** Not applicable. **Ethical approval:** Not applicable.

References

[1] Z. A. Khan, A. Beghdadi, F. A. Cheikh, M. Kaaniche, E. Pelanis, R. Palomar, Å. A. Fretland, B. Edwin, and O. J. Elle, "Towards a video quality assessment based framework for enhancement of laparoscopic videos," in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, vol. 11316, pp. 129–136, SPIE, 2020.

[2] H. H. Borate, P. A. Kara, B. Appina, and A. Simon, "A full-reference laparoscopic video quality assessment algorithm," in *Optics and Photonics for Information Processing XV*, vol. 11841, pp. 55–61, SPIE, 2021.

[3] S. Biswas and R. Palanisamy, "Quality evaluation of laparoscopic videos using the interdependency of luminance and texture maps," in *10th International Conference on Signal Processing and Integrated Networks*, pp. 103–108, IEEE, 2023.

[4] Z. A. Khan, A. Beghdadi, M. Kaaniche, and F. A. Cheikh, "Residual networks based distortion classification and ranking for laparoscopic image quality assessment," in *2020 IEEE International Conference on Image Processing*.

[5] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21.

[6] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[7] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[8] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.

[9] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.