On the Suitability of VMAF for Quality Assessment of Medical Videos: Medical Ultrasound & Wireless Capsule Endoscopy

Muhammad Arslan Usman, and Maria G. Martini*

Abstract—**With the rapid evolution in modern multimedia networks and systems, services such as telemedicine and tele-surgery are becoming more popular. Quality estimation and monitoring of medical videos is becoming important not only in the field of research, but also in real-time applications and services. The state-of-the-art video quality metric (VQM) called Video Multimethod Assessment Fusion (VMAF) is a promising solution for quality estimation of videos impaired by compression and scaling artifacts. The metric was developed by Netflix for entertainment video content and its good performance does not necessarily extend to medical video. This paper focuses on evaluating the performance of VMAF in the context of quality assessment (QA) for medical videos. We consider in this paper medical video compressed via High Efficiency Video Coding (HEVC) and refer in particular to medical ultrasound videos and wireless capsule endoscopy (WCE) videos for the performance estimation of VMAF. The correlation between the subjective scores of these two datasets and VMAF's quality estimates is studied and presented. The results show that VMAF outperforms other state-of-the-art VQMs in the context of WCE videos, but this is not the case for medical ultrasound videos.**

Index Terms—**High efficiency video coding, objective video quality assessment, subjective video quality assessment, medical ultrasound imagery, wireless capsule endoscopy.**

I. INTRODUCTION

Research, development and commercialization of multimedia systems, applications and services has witnessed an exponential growth in recent times. With billions of videos being streamed, shared, downloaded every day, it is crucial to maintain quality of service (QoS) and quality of experience (QoE) [1]. Especially for video-related applications and services the need for provision of quality of experience to the consumers becomes inevitable [1-3]. Telemedicine, image guided surgery, tele-surgery, etc., are becoming increasingly popular, as modern communication systems support high data rates hence allowing seamless delivery of videos to the end users [4]. Medical videos contain sensitive content which is of utmost importance to the clinicians and physicians. Large amounts of medical visual contents are created continuously on a daily basis for viewing and manipulation by medical professionals [5]. Video acquisition, processing, compression, transmission and display can result in inducing some artifacts, despite the fact that video capturing, and processing techniques are continuously evolving. Such artifacts in medical imagery may negatively impact the perception of medical professionals. As mentioned earlier, medical videos contain sensitive data, and their quality cannot be compromised as it might lead to medical errors [5].

High efficiency video coding (HEVC) has emerged as a promising solution for providing video compression without significantly reducing the video quality [6]. Bandwidth and storage limitations always prompt the service providers to adopt an efficient compression scheme. In medical videos, compression artifacts should not lead to medical errors as it can have dire consequences, in particular in the form of false diagnosis. Recent studies on HEVC compressed medical imagery, such as medical ultrasound and wireless capsule endoscopy (WCE) videos [3] [7-9], have shown that HEVC allows high amounts of compression without reducing the perceptual and diagnostic quality of the medical videos.

Objective video quality metrics (VQM) play an important role in the estimation of the perceptual quality of videos [2]. Though limited, several works have been done in the field of medical video quality assessment (VQA), which are discussed in the next Section. The Video Multimethod Assessment Fusion (VMAF) metric is a full-reference (FR) metric that estimates the quality based on compression and scaling artifacts [10]. Netflix, a leading USA-based media-services vendor, provides internet entertainment services such as TV series, documentaries, feature films, etc. Given the nature of Netflix's entertainment-oriented services, VMAF's application on medical videos has not been tested and verified before. Recent publications in the field of medical VQA have focused on the suitability of recent VQMs in the context of medical imagery. However, to the authors' knowledge VMAF has never been tested for its suitability for the estimation of the quality of compressed medical videos.

Date of re-submission: 22nd July 2019.

*Corresponding Author.

The authors are with Wireless Multimedia and Networking (WMN) Research Group, Faculty of Science, Engineering and Computing, Kingston University, KT1 2EE, London, United Kingdom. (Email: arslanusman@ieee.org, m.martini@kingston.ac.uk*).

Table I. Existing FR-VQMs considered as benchmark.

Quality metric	**Abbreviation**	**Description**
Peak Signal to Noise Ratio	PSNR	This metric is based on the calculation of the Mean Square Error (MSE).
Structural Similarity Index Metric [13]	SSIM	SSIM measures the quality of the video based on luminance, contrast and structural comparison between original and impaired videos.
Multi Scale SSIM [14]	MS-SSIM	MS-SSIM is an extension of SSIM with the same mathematical principles but estimates the quality of the image on multiple scales.
Visual Signal to Noise Ratio [15]	VSNR	Contrast thresholds are used to identify the impairments in the video sequences. All the impairments above these thresholds are mapped to represent the quality of the video sequences.
Information Fidelity Criterion [16]	IFC	Natural scene statistics (NSS) from the reference and impaired videos are combined to constitute a quality index.
Visual Information Fidelity [17]	VIF	Based on the Human Visual System (HVS), specific information is extracted from the reference video in wavelet domain. HVS refers to the information that can easily be extracted by the human brain from a video sequence. This same information is extracted from the impaired video sequence and then combined with the reference video information to measure the visual quality of the distorted video.
Pixel-based VIF [17]	VIFP	A simpler and less complex version of VIF is pixel-based VIF, which uses the same principles in the pixel domain to estimate the video quality.
Universal Quality Index [18]	UQI	UQI, like SSIM and MS-SSIM, measures the structural impairments in a video and then maps them to a model that predicts the quality.
Noise Quality Measure [19]	NQM	By considering the variation in contrast sensitivity, local luminance mean and contrast measures of the video sequence, this metric obtains a weighted signal to noise ratio measure between the reference and the processed video sequence.
Weighted Signal to Noise Ratio [19]	WSNR	WSNR, measured in dB scale, is calculated using the ratio between weighted signal power and noise power.
Video Quality Metric [20]	VQM $^{\text{NTIA}}$	Standardized by National Telecommunications and Information Administration (NTIA) USA, this metric estimates the quality based on seven different parameters from the reference and impaired videos.
Video Multimethod Assessment Fusion [10]	VMAF	As Netflix's video related services are based on the Transmission Control Protocol (TCP), the current version of VMAF estimates the video quality by considering only compression and scaling artifacts. The latest version of VMAF is based on support vector machine (SVM) regression which uses three features based on measurements from VIF and detail loss metric (DLM) [21] and temporal motion estimates. The motion estimation is done using a simple algorithm based on temporal difference of consecutive frames.

This paper focuses on studying the performance of VMAF in the context of medical videos. Two datasets are considered, namely medical ultrasound videos and wireless capsule endoscopy (WCE) videos. Both these datasets contain various types of ultrasound and WCE videos compressed via HEVC.

In this work the measurements of VMAF are fitted to subjective measurements of both the datasets in order to obtain a curve fitting model that produces best results. Correlations between subjective and objective measurements are also studied for comparison between the state-of-the-art quality metrics considered as benchmark and VMAF, and their results are discussed in detail.

In the following section we provide a survey of the state-of-the-art and the principal contributions of this work.

II. Background and Related Work

There have been several efforts in designing, standardizing and modelling VQMs specially designed for estimating the quality of medical videos. This section firstly discusses the state-of-the-art in VQA studies for medical videos and then discusses the contemporary FR objective VQMs.

A. *VQA in the context of Medical Videos & Images*

Quality assessment of medical imagery has been under limelight for quite a while and several relevant works have been published in the recent years. The authors in a recent survey [5] have detailed the subjective methods and findings of various medical image and video quality assessment studies. The survey encompasses 12 major studies that cover aspects of different medical imaging modalities. It includes three studies each about magnetic resonance imaging (MRI) and endoscopic imagery including WCE videos, one each about pathology imaging, heart imagery, ophthalmology videos and tele-surgery videos, and finally two about ultrasound videos.

B. *Objective Quality Metrics*

Objective video quality assessment is the least complex way of estimating the quality of visual content for various purposes, such as network optimization. Service providers employ objective video quality metrics to get automatic feedback of the video-related services, which consequently helps them to optimize the network. Such feedback is often used to prevent future encoding and transmission errors. For medical video-related services, such as telemedicine, this is very important. as preserving the diagnostic information is necessary.

Objective quality models can be classified into three major categories namely Full Reference (FR), Reduced Reference (RR) and No Reference (NR). The former two require full or partial reference of the original video, whereas the latter does not. FR methods are often used in cross-layer optimization [11], testing and validation of video compression methods, [6] etc. A detailed review of FR quality metrics can be found in [12]. A brief description of recent FR-VQMs, including Netflix's VMAF, is given in Table 1.

The FR metrics described in Table 1 are freely available online for research and academic purposes. These FR metrics are used in this paper for comparison purposes, using the recommended parameters recommended in the corresponding publications.

Inferring from the existing literature presented in this section and with the authors' best of knowledge, there has been no work so far which studies the performance of VMAF in the context of medical videos. The principal contributions of this paper are as follows:

- Performance of VMAF in quality estimation of HEVC compressed medical ultrasound videos and wireless capsule endoscopy videos.
- Presenting a curve fitting model for VMAF that produces best fit to the subjective DMOS for both video datasets.
- Comparison of VMAF with other state-of-the-art video quality metrics in terms of correlation between objective and subjective measurements.

The next section covers a concise description of both medical video datasets, i.e., with ultrasound and WCE videos.

III. Subjective Medical Video Datasets

The FR video quality metrics presented in Section II are designed to estimate the visual or perceptual quality of a video. These methods are not specifically designed for medical videos, so they are considered general purpose quality metrics. In order to assess the suitability of a video quality metric for specific visual content, it is important to conduct subjective experiments. Such measurements are used for evaluating the correlation with the objective VQMs' measurements.

In order to assess the performance and suitability of aforementioned VQMs in the context of medical videos, we have used two video datasets, described below.

A. *Dataset for Medical Ultrasound Videos [7]*

This dataset comprises nine different ultrasound videos, out of which three videos are related to heart and liver each, two to kidney and one to lungs. These nine videos have a spatial and temporal resolution of 640×416 and a frame rate of 25 frames per second (fps) respectively. With 100 frames in total for each video sequence, the total duration is 4 seconds. An example frame from each video with a brief description is available in [7]. These nine original videos were compressed at 8 different quantization parameter (QP) levels, ranging between 29 and 41, using the HEVC video encoder. A total of 72 HEVC compressed videos were evaluated by 4 medical experts and 15 non-experts.

The subjective measurements taken in this study use the double stimulus continuous quality (DSCQS) scale type-II methodology. The final measurements in this study are given in the form of differential mean opinion score (DMOS).

B. *Dataset for Wireless Capsule Endoscopy Videos [3]*

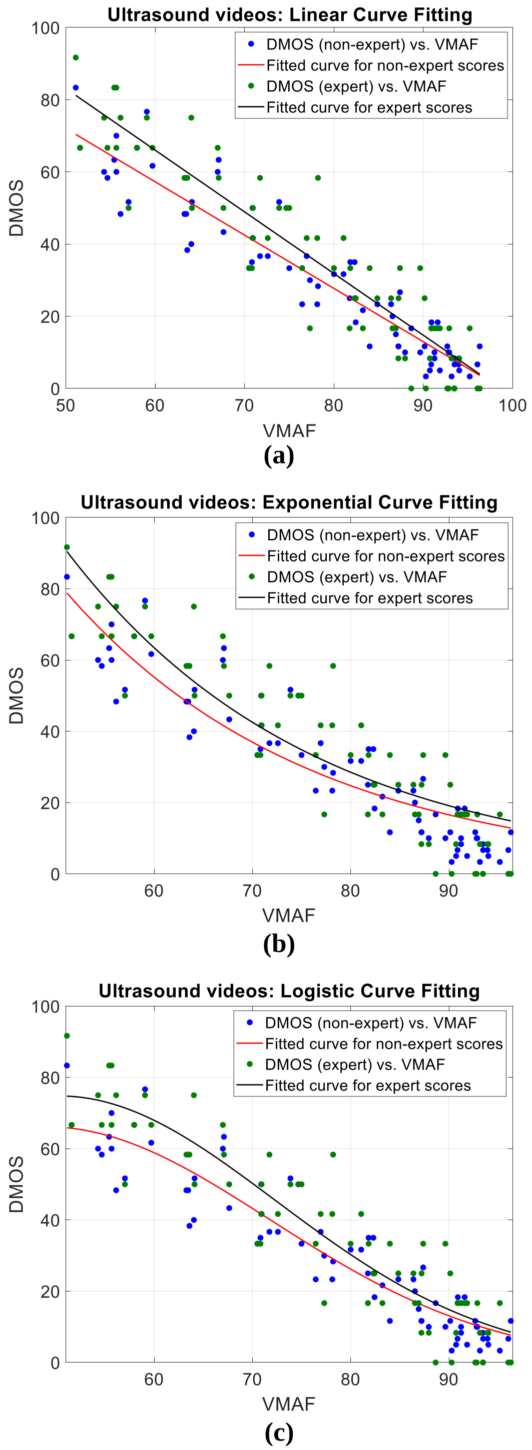
Wireless capsule endoscopy, or WCE, is a process in which a wireless capsule-shaped swallowable medical device is used to record imagery of the gastro-intestinal (GI) tract of living beings [22]. The information related to the WCE dataset provided in this subsection has been extracted from [3]. The WCE videos used in this study comprise ten different pathologies which are described in [3], along with the snapshot of each pathology. Each video in this dataset was compressed using the state-of-the-art HEVC encoder at eight different compression levels. Similar to the ultrasound dataset, the QP range was kept between 27 and 41, with a step size of 2. With a spatial and temporal resolution of 320×320 and 3 fps respectively, each video is 10 seconds in duration. The total number of videos in this dataset is 90, with 10 original videos compressed at eight different compression levels resulting in 80 HEVC compressed videos.

The scoring method used for the subjective measurements of this dataset was the same as the Ultrasound videos dataset, i.e., DSCQS type-II. The results, which were collected from 6 experts and 18 non-experts, are in the form of DMOS.

IV. Results and Discussion

The subjective tests for the aforementioned datasets and their corresponding results are thoroughly presented in [3] [7]. In this section, we have used the subjective measurements in the form of DMOS from these two datasets to evaluate the performance of VMAF and of the other FR-VQMs considered for comparison.

In order to quantify the relationship between measurements from FR-VQMs and the subjective measurements, the objective measurements are fitted to curve fitting models. The curve



(a)

(b)

(c)

Fig. 1. Results for fitting the VMAF measurements of ultrasound videos to subjective DMOS of both experts and non-experts. (a) Linear model, (b) Exponential model, (c) Logistic model.

fitting in the next subsections is done for both experts' and non-experts' DMOS and the results for both datasets are discussed separately.

A. Results for the Ultrasound Videos Dataset

For the performance evaluation of VMAF in the context of ultrasound videos, we have considered exponential, linear and logistic curve fitting approaches, where all three exhibit monotonic curves. The results are reported in terms of coefficient of determination (R^2), adjusted coefficient of determination (Adj.R^2) and root mean square error (RMSE). These performance metrics are calculated for the aforementioned three types of curve fitting approaches.

First, we have used the simplest of curve-fitting approaches i.e. linear curve fitting and the results are shown in Fig. 1a. It can be observed that the VMAF measurements exhibit a good fit to the subjective DMOS of both experts and non-experts. The mathematical representation of linear fitting is given in eq. (1), where Z_j and Z_j' represent mean score before and after fitting for the j$^{\text{th}}$ video sequence respectively. The parameters β_1 and β_2 are estimated using the *nlinfit* tool in MATLAB.

$$Z_j' = \beta_1 Z_j + \beta_2 \, . \qquad (1)$$

The results in Fig. 1b show that using the exponential approach, VMAF measurements for the ultrasound videos exhibit a better fit to the subjective DMOS of both experts and non-experts. The mathematical expression for exponential curve fitting is given as follows.

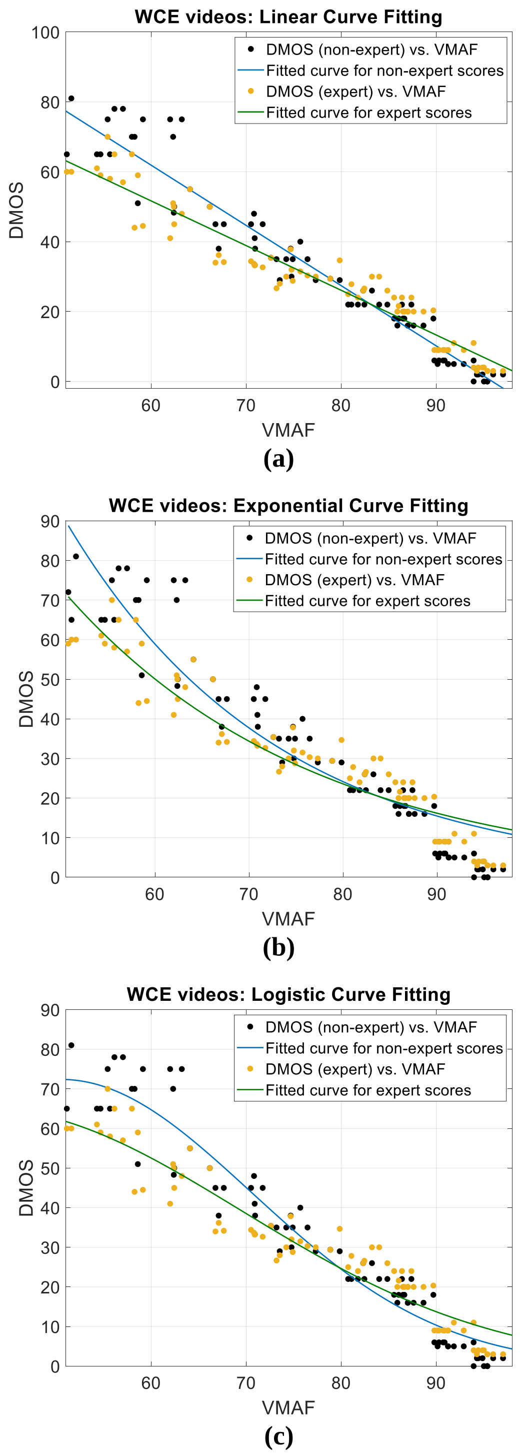
$$Z_j' = \beta_1 \exp(Z_j \beta_2). \qquad (2)$$

Finally, we have used four-parameter logistic curve fitting, as it is one of the most common monotonic approaches used in the context of VQA. The authors in [7] also use the same approach for fitting the objective measurements to the subjective DMOS of ultrasound videos. Fig. 1c shows that the VMAF measurements for the ultrasound videos exhibit an excellent fit to the DMOS measurements using the logistic model. The mathematical expression for this model is given in eq. (3) and the four parameters ($\beta_1, \beta_2, \beta_3$ & β_4) are estimated the same way i.e. using the *nlinfit* tool in MATLAB.

$$Z_j' = \beta_2 + \frac{\beta_1 - \beta_2}{1 + exp\left(-\left(\frac{Z_j - \beta_3}{|\beta_4|}\right)\right)} \, . \qquad (3)$$

Table 2 contains the numerical results for all the three curve fitting approaches for both expert and non-expert DMOS. It can be observed that in terms of both R^2 and RMSE, VMAF shows best performance when Linear fitting is used. The exponential and logistic curve fitting approaches are comparable in terms of R^2, but in terms of RMSE the latter performs better for both expert and non-expert DMOS. So, it can be inferred that for ultrasound videos, linear fitting produces the best fit of VMAF w.r.t.. expert and non-expert DMOS. The RMSE and R^2 results for experts' DMOS are lower as compared to non-experts, the reason being that the former assesses the quality of medical videos in terms of diagnostic quality only.

Furthermore, we have compared the performance of VMAF with other VQMs in terms of correlation. In Table 3, the results for Pearson's linear correlation coefficient (PLCC) and Spearman's rank order correlation coefficient (SROCC) are presented. The presented correlation values have been calculated between the objective VQMs' measurements and DMOS from the ultrasound videos dataset. For experts' DMOS it can be observed that, in terms of PLCC, VMAF shows better performance as compared to MS-SSIM, VSNR, NQM, VIFP



(a)

(b)

(c)

Fig. 2. Results for fitting the VMAF measurements of WCE videos to subjective DMOS of both experts and non-experts. (a) Linear model, (b) Exponential model, (c) Logistic model.

and IFC, whereas for SROCC it only performs better than two metrics i.e. MS-SSIM and IFC. For non-experts' DMOS, VMAF performs better than PSNR, SSIM, MS-SSIM, VSNR, WSNR, VIFP and IFC in terms of PLCC, but for SROCC its performance is only better than MS-SSIM, VIFP and IFC. Overall, for ultrasound videos, UQI and VIF are the best performing metrics in terms of PLCC and SROCC respectively, for both experts' and non-experts' DMOS.

In [7], the authors have reported a compression threshold of $QP = 35$ in terms of maximum allowed compression for diagnostically acceptable quality for ultrasound videos. This threshold was suggested based on results from the experts' subjective scores. In Table 4, objective metrics' results, including VMAF, are reported that correspond to each ultrasound video compressed at $QP = 35$. The nomenclature in Table 4 for each ultrasound video has been taken from [7].

Ultrasound videos are non-conventional videos as compared to other medical videos, such as WCE and entertainment videos. The support vector classifier (SVC) in the VMAF metric is trained on general videos as NETFLIX's target audience is from the entertainment domain.

The next section reports the evaluation of VMAF for the second dataset i.e. WCE videos dataset.

B. Results for the WCE Videos Dataset

We have used the same three curve fitting approaches for the WCE dataset, and the results are provided in Fig. 2a, 2b and 2c for linear, exponential and logistic fitting respectively. Observing the results for all the three monotonic fits, it can be observed that they exhibit a good fit, yet almost the same, for both experts' and non-experts' DMOS. The curve fitting results in terms of RMSE and R^2 for the WCE dataset are also presented in Table 2. It can be observed that, compared to the Ultrasound videos dataset, the performance of VMAF is much better for the WCE videos dataset in terms of R^2 and RMSE values. In terms of a better fit for experts' DMOS, the results follow a different trend as compared to ultrasound videos. It can be seen that the R^2 values are highest for linear fitting, followed by the logistic fit and exponential fit. All the three fits exhibit approximately the same results, with negligible difference at third decimal place, in terms of R^2. Further for the experts' DMOS, in terms of RMSE, VMAF shows better performance using the logistic fit as compared to linear fit and shows comparable performance to the exponential fit. For the non-experts' DMOS, the exponential fit shows the best results, in terms of both R^2 and RMSE, followed by logistic and linear fit respectively.

Like ultrasound videos, the diagnostically acceptable quality for WCE videos was suggested $QP = 35$ and 37 based on subjective scores of experts and non-experts respectively. In this paper we have mainly focused on expert opinion, so in

Table II. Results for fitting the VMAF measurements to the subjective DMOS.

Dataset	**Category**	**Exponential**			**Linear**			**Logistic**		
		R^2	**Adj. R^2**	**RMSE**	**R^2**	**Adj. R^2**	**RMSE**	**R^2**	**Adj. R^2**	**RMSE**
Ultrasound Videos	**Expert**	0.8032	0.8004	11.1736	0.8620	**0.8601**	**9.3557**	0.8544	0.8502	9.680
	Non-Expert	0.8334	0.8310	8.7630	0.8862	**0.8846**	**7.242**	0.8791	0.8756	7.5199
WCE Videos	**Expert**	0.9214	0.9204	3.8859	0.9267	**0.9258**	4.9820	0.9268	0.9239	**3.7994**
	Non-Expert	0.9501	**0.9494**	**3.0967**	0.9433	0.9426	5.8556	0.9501	0.9481	3.1370

Table III. Comparison of VMAF with other FR-VQMs

Dataset	Scores	CC	PSNR	SSIM	MS-SSIM	VSNR	WSNR	NQM	UQI	VIF	VIFP	IFC	VQM^{NTIA}	VMAF
Ultra-sound	**Experts**	**PLCC**	0.9109	0.9264	0.8570	0.8925	0.9123	0.8961	**0.9292**	0.9258	0.8887	0.8644	0.8080	0.9056
		SROCC	0.9331	0.9375	0.8907	0.9139	0.9251	0.9090	0.9251	**0.9382**	0.8997	0.8926	0.8368	0.8941
	Non-experts	**PLCC**	0.8896	0.9208	0.8668	0.8888	0.9173	0.9233	**0.9520**	0.9431	0.8796	0.8446	0.8146	0.9220
		SROCC	0.9280	0.9383	0.8899	0.9277	0.9354	0.9464	0.9495	**0.9663**	0.9047	0.8906	0.8606	0.9186
WCE	**Experts**	**PLCC**	0.8039	0.6840	0.8366	0.6055	0.8010	0.7158	0.8701	0.9016	0.8955	0.8844	0.7764	**0.9627**
		SROCC	0.8611	0.8063	0.9127	0.6571	0.8709	0.8257	0.8930	0.9424	0.9263	0.9482	0.8426	**0.9763**
	Non-experts	**PLCC**	0.8257	0.7232	0.8696	0.6204	0.7963	0.7371	0.8909	0.9238	0.9227	0.9020	0.7578	**0.9712**
		SROCC	0.8642	0.8129	0.9247	0.6474	0.8774	0.8311	0.9061	0.9533	0.9408	0.9525	0.8402	**0.9796**

Table IV. Quality Values for FR-VQMs corresponding to acceptable diagnostic quality in terms of compression in WCE videos

Video Sequences	PSNR	SSIM	MS-SSIM	VSNR	WSNR	NQM	UQI	VIF	VIFP	IFC	VQM^{NTIA}	VMAF
Angiodysplasia	30.7381	0.8585	0.9351	20.0233	27.8509	18.4914	0.6164	0.2707	0.3889	1.4493	1.8433	77.308
Ascaris	34.9131	0.9141	0.9579	26.3202	30.7488	21.2988	0.6269	0.3102	0.4753	1.2059	1.2615	76.4479
Crohn's Disease	35.7274	0.9153	0.9567	27.6202	32.6963	22.5215	0.592	0.3177	0.4551	1.2372	1.1593	75.6877
Diverticulum	35.2772	0.9129	0.9538	27.2796	30.8014	22.2541	0.5661	0.2839	0.4391	1.0523	1.2454	79.8377
Phlebectasia	34.7786	0.8952	0.943	26.0166	31.9825	21.1328	0.5317	0.2506	0.4018	0.9719	1.2192	73.5249
Polyp	36.3554	0.9363	0.9643	32.4502	29.6683	23.0757	04806	0.3336	0.4607	1.0953	1.3219	78.8445
Stenosis	35.3696	0.9138	0.9515	28.7107	30.2745	21.5594	0.4953	0.2771	0.4102	0.9817	1.2306	74.763
Subepithelial Tumor	36.4692	0.926	0.9572	27.9799	31.4753	21.6927	0.5627	0.3209	0.4454	1.14	1.0842	74.2097
Tumor	34.0502	0.8871	0.9398	22.6127	31.3318	19.6339	0.5621	0.2758	0.4071	1.1649	1.3467	74.8874
Xanthoma	35.2061	0.8975	0.9421	23.5172	32.0486	19.8628	0.495	0.2625	0.3916	1.0222	1.2149	73.1839

Table V. Quality Values for FR-VQMs corresponding to acceptable diagnostic quality in terms of compression in Ultrasound videos

Video Sequences	PSNR	SSIM	MS-SSIM	VSNR	WSNR	NQM	UQI	VIF	VIFP	IFC	VQM^{NTIA}	VMAF
Sequence A	33.6696	0.9030	0.9432	39.1150	26.3375	27.7196	0.7605	0.3785	0.4026	1.3070	1.3921	78.2096
Sequence B	34.6696	0.9340	0.9555	41.5616	28.5047	29.1442	0.7932	0.4218	0.4469	1.2471	1.3962	78.1267
Sequence C	34.4605	0.9197	0.9575	40.9681	29.2536	28.9021	0.7850	0.4063	0.4319	1.2178	1.2711	77.3104
Sequence D	33.8895	0.9197	0.9549	40.4798	29.1883	30.7338	0.7808	0.4235	0.4639	1.5282	1.2881	81.8238
Sequence E	34.1383	0.9122	0.9443	37.5422	24.3815	25.0479	0.7630	0.3791	0.3889	1.2058	1.3244	73.8655
Sequence F	33.6477	0.9117	0.9416	40.3240	27.9050	29.0924	0.7924	0.4079	0..4349	1.4746	1.3864	79.9976
Sequence G	34.9024	0.9342	0.9503	39.2744	26.6547	27.3438	0.7815	0.4443	0.4504	1.3296	1.2035	76.9455
Sequence H	33.1649	0.8963	0.9402	37.8752	27.7311	26.9436	0.7395	0.3706	0.3797	1.2958	1.3944	74.9981
Sequence I	34.2645	0.8974	0.9528	39.2703	26.1668	25.9475	0.6932	0.3389	0.3425	0.9752	1.3012	76.4415

Table 5, values for objective metrics are provided that correspond to WCE videos compressed at $QP = 35$. The nomenclature for WCE videos in Table 5 is taken from [3].

Comparing the performance of VMAF with other video quality metrics, it can be observed in Table 3 that VMAF outperforms all other VQMs in terms of both PLCC and SROCC. The authors in [3] have concluded VIF to be the best performing metric according to their study, but it can be seen that VMAF outperforms VIF with a clear margin for both experts' and non-experts' DMOS.

V. Conclusion

In this paper, we evaluated the performance of the state-of-the-art video quality objective metric VMAF in the context of medical videos. The metric's ability to correctly estimate the quality of two types of medical videos, ultrasound and WCE, was tested and compared to other contemporary metrics. The VMAF measurements were fitted to the subjective DMOS of expert and non-expert observers using exponential, linear and logistic curve fitting models. The linear fitting model exhibited the best fit in terms of R^2 and RMSE for the ultrasound videos for both experts' and non-experts' DMOS. In case of WCE videos, for the experts' DMOS the linear model exhibited the best fit only in terms of R^2, but in terms of RMSE, the logistic fit exhibited the best fit. For the non-experts' DMOS the exponential model exhibited the best fit in terms of both R^2 and RMSE.

Furthermore, the presented results indicated different outcomes for the two considered video datasets in terms of PLCC and SROCC, as VMAF outperformed all other metrics for HEVC compressed WCE videos, but for ultrasound videos this is not the case. The results for ultrasound videos are substantially different from those for WCE videos. According to authors' understanding, the reason appears to be that ultrasound videos are different in terms of the capturing process. Ultrasound videos are produced through multiple scans, called sonograms, and the sound waves are used to create an image of an internal organ of the body. Hence, conventional image capturing processes, such as the CMOS camera in WCE, are not used in ultrasounds which makes the video not well represented / assessed by a quality metric mainly trained for natural videos.

VI. Future Work

The performance of VMAF showed very good results in terms of quality estimation for wireless capsule endoscopy videos but the same was not observed for ultrasound videos. This leaves room for improvement in VMAF's performance for medical videos, specifically ultrasound videos. Based on the reasons mentioned in conclusion section, VMAF can be trained for a large dataset of ultrasound videos in order to improve its quality estimation.

Furthermore, VMAF can be tested for other types of medical imagery, as mentioned in Section II of this paper.

Acknowledgement

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme through the Marie Skłodowska-Curie Grant under Agreement 643072.

References

[1] M. A. Usman, M. R. Usman, and S. Y. Shin, "Exploiting the Spatio-Temporal Attributes of HD Videos: A Bandwidth Efficient Approach," in *IEEE Trans. on Circuits and Systems on Video Tech.*, vol. 28 (9), pp. 2418-2422, Jul. 2018.

[2] M. A. Usman, M. R. Usman, and S. Y. Shin, "A Novel No-Reference Metric for Estimating the Impact of Frame Freezing Artifacts on Perceptual Quality of Streamed Videos", in *IEEE Trans. on Multimedia*, Vol. 20 (9), pp. 2344–2359 Sept. 2018.

[3] M. A. Usman, M. R. Usman, and S. Y. Shin, "Quality assessment for wireless capsule endoscopy videos compressed via HEVC: From diagnostic quality to visual perception," *Comput. Biol. Med.*, vol. 91, pp. 112–134, Dec. 2017.

[4] M. G. Martini, "Wireless broadband multimedia health services: Current status and emerging concepts," in Proc. *IEEE 19th Int. Symp. Personal, Indoor Mobile Radio Commun*., Nantes, France, Sep. 2008, pp. 1–6.

[5] Lévêque, L. et al., On the Subjective Assessment of the Perceived Quality of Medical Images and Videos. In *IEEE Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018.

[6] J. Ohm, G.J. Sullivan, H. Schwarz, T.K. Tan, T. Wiegand, Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC), in *IEEE Trans. Circuits Syst. Video Technol*. Vol. 22 (12) pp. 1669–1684, Dec. 2012.

[7] M. Razaak, M.G. Martini, K. Savino, A study on quality assessment for medical ultrasound video compressed via HEVC, *IEEE J. Biomed. Health Inf.* 18 (5), pp. 1552–1559, 2014.

[8] Panayides, A. S., et al., An effective ultrasound video communication system using despeckle filtering and HEVC. *IEEE journal of biomedical and health informatics*, vol. 19, no. 2, pp: 668-676, 2014.

[9] Fernández, D.G., et al., HEVC optimizations for medical environments. In *Sensing and Analysis Technologies for Biomedical and Cognitive Applications*, Vol. 9871, p. 98710B, International Society for Optics and Photonics, May 2016.

[10] Netflix. VMAF - Video Multi-Method Assessment Fusion. https://github.com/Netflix/vmaf. [Online: accessed 12-Dec-2018].

[11] Zhao, P., Liu, Y., Liu, J., Argyriou, A. and Ci, S., 2016. SSIM-based error-resilient cross-layer optimization for wireless video streaming. *Sig. Processing: Image Comm., 40,* pp.36-51.

[12] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furht and O. Marques, Eds. CRC Press, 2003.

[13] Z.Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[14] W., Zhou, E. P. Simoncelli, and A. C. Bovik. "Multiscale structural similarity for image quality assessment." *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*. Vol. 2. Ieee, 2003.

[15] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process*., vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[16] Sheikh, Hamid Rahim, Alan Conrad Bovik, and Gustavo De Veciana. "An information fidelity criterion for image quality assessment using natural scene statistics." *IEEE Transactions on Image Processing,* 14.12 (2005): 2117-2128.

[17] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, pp. 127–135, 2013.

[18] Z.Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Letters,* vol. 9, no. 3, pp. 81–84, Mar. 2002.

[19] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process*., vol. 9, no. 4, pp. 636–650, Apr. 2000.

[20] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast*., vol. 50, no. 3, pp. 312–322, Sep. 2004.

[21] S. Li, F. Zhang, L. Ma, and K. Ngan, "Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments," *IEEE Transactions on Multimedia,* vol. 13, no. 5, pp. 935–949, Oct. 2011.

[22] M. A. Usman, G. B. Satrya, M. R. Usman, and S. Y. Shin, Detection of small colon bleeding in wireless capsule endoscopy videos, *Computerized Medical Imaging and Graphics*, Vol. 54, pp. 16–26, 2016.