Visual quality assessment of H.264/AVC compressed laparoscopic video

Asli Kumcu[a], Klaas Bombeke[b], Heng Chen[c], Ljubomir Jovanov[a], Ljiljana Platiša[a], Hiep Luong[a], Jan Van Looy[b], Yves Van Nieuwenhove[d], Peter Schelkens[c], Wilfried Philips[a]

[a]iMinds-IPI-Ghent University, Ghent, Belgium;
[b]iMinds-MICT-Ghent University, Ghent, Belgium;
[c]iMinds-ETRO-Vrije Universiteit Brussel, Brussels, Belgium;
[d]Department of Gastrointestinal Surgery, Ghent University Hospital, Ghent, Belgium;

ABSTRACT

The digital revolution has reached hospital operating rooms, giving rise to new opportunities such as tele-surgery and tele-collaboration. Applications such as minimally invasive and robotic surgery generate large video streams that demand gigabytes of storage and transmission capacity. While lossy data compression can offer large size reduction, high compression levels may significantly reduce image quality. In this study we assess the quality of compressed laparoscopic video using a subjective evaluation study and three objective measures. Test sequences were full High-Definition videos captures of four laparoscopic surgery procedures acquired on two camera types. Raw sequences were processed with H.264/AVC IPPP-CBR at four compression levels (19.5, 5.5, 2.8, and 1.8 Mbps). 16 non-experts and 9 laparoscopic surgeons evaluated the subjective quality and suitability for surgery (surgeons only) using Single Stimulus Continuous Quality Evaluation methodology. VQM, HDR-VDP-2, and PSNR objective measures were evaluated. The results suggest that laparoscopic video may be lossy compressed approximately 30 to 100 times (19.5 to 5.5 Mbps) without sacrificing perceived image quality, potentially enabling real-time streaming of surgical procedures even over wireless networks. Surgeons were sensitive to content but had large variances in quality scores, whereas non-experts judged all scenes similarly and over-estimated the quality of some sequences. There was high correlation between surgeons' scores for quality and "suitability for surgery". The objective measures had moderate to high correlation with subjective scores, especially when analyzed separately by camera type. Future studies should evaluate surgeons' task performance to determine the clinical implications of conducting surgery with lossy compressed video.

Keywords: video compression, quality assessment, H.264/AVC, laparoscopy, telesurgery, telemedicine

1. INTRODUCTION

Laparoscopic surgery is a type of minimally invasive surgery, which reduces trauma to the patient compared to open surgery. The laparoscopic surgeon uses specialized instruments and an optical camera (endoscope) inserted into the body through small incisions to conduct the surgery; the surgical field is viewed indirectly via a display. A video recording of the surgery can be stored or transferred within and outside the hospital, for example for teaching or tele-collaboration. However, these systems produce high-resolution long-duration video streams that demand gigabytes of storage and transmission capacity. A camera acquiring video in raw RGB with 8 bits per channel in high definition (HD) resolution (1920x1080 pixels), at 25 frames per second (FPS), will generate approximately 148 MB of raw data every second. A single 60-minute procedure will require around 520 GB of storage space if stored in raw format.

Compression can be used to reduce video size and improve data storage and transmission efficiency. While lossless compression may be preferred for extremely critical data, lossy compression offers greater size reduction. At low levels of compression, image quality is not at all, or very little, affected by lossy compression; however, excessively high compression levels can distort the scene and reduce image quality to clinically unacceptable levels. Current camera and hospital IT systems may reduce video size either by decreasing the video resolution or by applying standard compression; while this improves video portability, it may limit its clinical utility.

Send correspondence to Asli Kumcu, E-mail: asli.kumcu@telin.ugent.be

As the current state-of-the-art video compression standard has moved beyond MPEG to H.264/AVC,[1] it is important to benchmark this new standard for medical applications. However, the authors are not aware of any existing studies examining the clinical quality impact of H.264/AVC lossy compression on full HD resolution laparoscopic surgery video. Existing studies have examined tele-endoscopy or robotic surgery without compression, MPEG compression variants, or lower resolution H.264 sequences. One study that examined MPEG-2 compression in four robotic surgery scenes using the Double-Stimuli Continuous Quality Scale (DSCQS) protocol concluded that sequences could be compressed to 3.2 megabits per second (Mbps) - a compression ratio of 90:1 - with no loss in perceived quality.[2] A study that evaluated four codecs (MJPEG, MPEG-1, MPEG-2 [4:2:0], MPEG-2 [4:2:2]) with 40 endoscopic sequences using overall image quality and usability for diagnosis found that only sequences compressed with MPEG-2 [4:2:2] at 40 Mbps were indistinguishable from the uncompressed sequence, and in addition MPEG-2 [4:2:0] at or above 8 Mbps and MJPEG at 15 Mb/s were "more or less acceptable usability for diagnosis".[3] Another study that examined the effect of four codecs (M-JPEG2000, MPEG-2, MPEG-4, H.264) on 4 bronchoscopy sequences with 256 x 256 resolution using a modified Stimulus Comparison Adjectival Categorical Judgement (SCACJ) protocol found that H.264 had the best overall quality and sequences could be compressed up to 1.34 Mbps (for dynamic video) and 0.84 Mbps (for static video) without any loss in perceived quality.[4]

In this study, we assess the quality of full HD resolution laparoscopic video compressed with H.264/AVC optimized for low latency using a subjective evaluation study and state-of-the-art objective measures. The primary goals of this work are to determine which compression levels generate clinically acceptable sequences, the effect of scene (content) on quality judgments, and the effect of the subjective task. In addition, we aim to determine whether quality judgments are independent of expertise by comparing the quality ratings of non-experts and surgeons. If non-experts judge the quality of different compression levels the same as surgeons, the former could be used as surrogates for the latter. Finally, we aim to evaluate the performance of objective quality assessment (QA) measures in predicting the quality judgments of surgeons. Since it can be difficult to recruit a sufficient number of experts for subjective studies, the use of either non-expert subjects, objective measures, or both, could potentially accelerate research and development time.

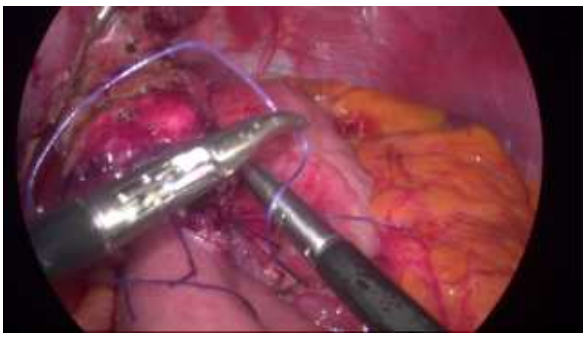
The methodology of the subjective and objective studies is explained in section 2, results are presented in section 3 with a discussion of the implications in section 4, and concluding remarks in section 5.

2. METHODS & MATERIALS

2.1 Stimuli

The four test sequences, shown in Fig. 1, were captured during abdominal laparoscopic procedures at a local hospital. Two procedures used a standard optical camera system (scenes "A" and "B" in Figs. 1a and 1b) and two used a chip-on-tip digital camera system (scenes "C" and "D" in Figs. 1c and 1d). All sequences were acquired as full HD (1920x1080 pixel resolution) interlaced-scan video in uncompressed RGB format with a bit-depth of 8 bits per component at 25 FPS. Each frame was converted from RGB to YCbCr 4:4:4,[5] followed by chroma subsampling to YCbCr 4:2:0. A 10-second segment was extracted from each of the four surgeries (*scenes*) to generate one uncompressed *reference* sequence per scene. Each reference video was then compressed at four H.264/AVC[6] compression levels with the x264[7] software, optimizing for low-latency (*IPPP-CBR*), using the following parameters: profile high, preset medium, tune zerolatency, vbv-maxrate *kbps*, vbv-bufsize *kbps*, intra-refresh, nal-hrd cbr, keyint 25, slices 1, min-keyint 25, no-scenecut, ref 1, bitrate *kbps*, ipratio 1.20, chroma-qp-offset 0, partitions all, direct auto, me umh, merange 128, weightp 0, psy, no-fast-pskip, 8x8dct, level 41, bframes 0, force-cfr, no-mbtree, sync-lookahead 0, sliced-threads, rc-lookahead 0, ratetol 0.01 (where *kbps* is the bit-rate in kilobit per second (Kbps)). All sequences were subsequently deinterlaced with YADIF[8] (mode 1, double framerate, with temporal and spatial interlacing check).

Sequences were compressed at 19.5, 5.5, 2.8, and 1.8 megabit per second (Mbps) , corresponding to compression ratios of approximately 31, 111, 214, and 336. The compression ratio was calculated as the ratio of the size on disk of the reference (YCbCr 4:2:0) and test (H.264/AVC compressed) videos prior to deinterlacing; note that the conversion from raw RGB to YCbCr 4:2:0 results in a 50% reduction of the size of the video sequence; therefore, compression ratio values should be doubled to compare disk savings with respect to raw RGB video.



(a) scene A

(b) scene B

(c) scene C

(d) scene D

Figure 1. Examples frames from the four reference laparoscopic video sequences.

Compression levels were selected with a pilot study (not reported here) to ensure that the quality differences between the four compression levels were approximately equal. The smallest level (20 Mbps) was chosen to be sufficiently equivalent in quality to the reference sequence, while the largest compression level (1.8 Mbps) was sufficiently degraded in quality while still allowing some visualization of vasculature.

A total of 20 stimuli were generated (4 scenes each consisting of 1 reference and 4 compressed sequences).

2.2 Participants

Participants from two subject groups were recruited for the study: laparoscopic surgeons (experts) and non-experts.

The nine experts were laparoscopic surgeons specialized in abdominal surgery from Ghent University Hospital. Five were first to fourth year residents of which two had 2-5 years experience with laparoscopic surgery and three had 6 or more years experience. The other four subjects were staff surgeons with more than 6 years experience. There were six males and three females; most did not indicate their age. Seven of the surgeons rated all sequences for four scenes, one rated all sequences for scenes A and C only, and one rated all sequences for scenes B and D only.

The sixteen non-experts were doctoral students and post-doctoral researchers with experience in image processing, analysis, and restoration. Twelve indicated a high level of expertise in image processing and eleven indicated a moderate to high level of experience with subjective image quality assessment. All had at least one previous experience conducting quality assessment studies with natural scene videos. There were thirteen males and three females, between the ages of 25 and 37. All non-experts rated all sequences for all four scenes.

All subjects were compensated for their time with gift certificates.

2.3 Subjective quality assessment

Subjects received an explanation of the experimental protocol and two vision tests. They were screened for visual defects with a Snellen chart visual acuity test (non-experts only) and a digital Ishihara Compatible color vision test.

Video quality was evaluated using the Single Stimulus Continuous Quality Evaluation (SSCQE)[9] methodology. The twenty sequences were presented in a randomized fashion one at a time. Both subject groups were familiar with the range of scenes and compression levels due to participation in two paired-comparison quality assessment studies using the same test sequences (not reported here) prior to the SS experiment. Both non-experts and surgeons were asked to rate the overall quality ("Quality") of each sequence using a continuous scale from 0 (Poor) to 100% (Excellent quality). In order to evaluate the effect of the subjective task, surgeons were also asked to rate whether they thought the sequences were suitable for use during surgery ("Suitability for surgery") on a continuous scale from 0 (Unusable) to 100% (Excellent).

Two 24" full High Definition resolution (1920x1200 pixels) surgical displays (MDSC-2124, Barco, Kortrijk, Belgium) were used to display the test interface and the sequences. The displays were color-matched and set at 100% luminance (400 cd/m^2), Gamma display function (2.2 gamma), and 6500K color temperature, with noise reduction and sharpness post-processing disabled. Viewers were seated approximately 90 cm from the displays. Experiments were conducted in low room illumination with indirect lighting (<200 lux).

2.4 Objective quality assessment

Objective measures are computer algorithms that aim to predict the quality judgments of human subjects. They are frequently used in consumer media applications, such as television and internet video, as surrogates for human quality judgments. Objective measures are typically classified into three categories: full-reference, in which the post-processed video is directly compared to the original; reduced-reference, in which a subset of features extracted from each video are compared; and no-reference, in which a single video is analyzed independently. In this study, two state-of-the-art measures were evaluated: the reduced-reference VQM v1.4 General Model[10] (National Telecommunications and Information Administration (NTIA), United States), and the frame-based, full-reference HDR-VDP-2.1.3[11] mean-opinion-score prediction model; peak signal-to-noise-ratio[12] (PSNR) was also evaluated. Full- and reduced-reference measures generate one score per *degraded* sequence indicating either the quality loss compared to the reference, or the overall quality .

NTIA's VQM is a general purpose video quality model that has been standardized by the American National Standards Institute (ANSI) and included in International Telecommunication Union (ITU) recommendations.[10] This model was chosen for inclusion in this study because it is publicly available, computationally fast, and shown to perform well in a variety of Video Quality Experts Group (VQEG) tests. Video quality is predicted by comparing spatiotemporal features extracted from the reference and degraded sequences; perceptual factors, as well as color and temporal information, are taken into account. VQM generates one perceived impairment score (VQM *index*) per degraded sequence with higher scores indicating worse impairment (quality). No calibration metrics were used as it was assumed that the original and processed videos matched exactly with respect to: spatial scaling and spatial shift, temporal shift, and luminance gain offset. Videos were converted to RGB24 before processing.

HDR-VDP-2 is a "high dynamic range visual difference predictor" built using a realistic human visual model – including the eye's contrast sensitivity function (CSF), masking effects, response of receptors to a variety of stimuli, and neural noise – and takes into account viewing conditions such as display luminance/contrast and viewing distance.[11] The measure does not take color or temporal aspects into account. This model was chosen because of its ability to realistically model the human visual system, at least in the spatial domain. The mean of the HDR-VDP-2 quality scores over all frames was used as the quality score (HDR-VDP-2 Q_{MOS}%) for each degraded sequence, with higher scores indicating better quality. Only the luma (Y') channel of the sequence was used to compute the HDR-VDP-2 scores.

Despite poor correlation of PSNR with human perception, it was included in this experiment to facilitate comparisons with other studies. The mean of the PSNR quality scores over all frames was used as the quality score (PSNR dB) for each degraded sequence, with higher scores indicating better quality.

2.5 Data analysis

Significance testing for the effects of compression and content (*within-subjects* testing) was conducted with the repeated-measures Friedman test (with post-hoc testing corrected for multiple comparisons,[13] α <0.05). Non-parametric testing was preferred due to few sample points (subjects) and non-normally distributed scores per sequence. The effect of compression level was analyzed on each scene separately; the effect of content was analyzed for each compression level separately. For expert scores only, correlation between quality and suitability for surgery scores was evaluated with Spearman and Kendall's tau rank correlation coefficients. The effect of expertise (*between-subjects* testing) for each scene and compression level was tested with the Wilcoxon rank-sum test (α <0.05), while the correlation of the median scores was compared using Spearman and Kendall's tau rank correlation coefficients. Spearman correlation is a measure of how far the ranks of the scores differ, whereas Kendall's tau gives the probability of agreement. Analysis of subjective scores was conducted using the R statistical package.[14] The clinically unacceptable levels of compression were defined as compression levels whose scores were statistically significantly different from the reference (uncompressed) sequence.

Outlier testing of subjective quality scores was conducted as given in the ITU-R recommendations.[9] Quality scores were converted to subjective difference median opinion scores ($DMdnOS$) by taking the difference in median scores between the reference and compressed sequences, in order to compare with the objective measures. Note that this study uses the median, rather than the mean scores, due to the non-normal distribution of the subjective scores. $DMdnOS$ were fit to objective scores with a nonlinear function that was then used to predict the subjective scores ($DMdnOS_p$). The logistic regression method outlined in ITU-R recommendations[9] was conducted in Matlab (The MathWorks, Inc., Natick, Massachusetts, United States) using the following fitting function to optimize the estimation of the β coefficients:

$$DMdnOS = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-(X-\beta_3)/|\beta_4|}}, \tag{1}$$

where X is a vector of objective scores containing one value per stimulus and $DMdnOS$ is the vector of corresponding subjective scores. Predicted subjective scores ($DMdnOS_p$) were then computed using Eq. 1 with the given X and computed β_{1-4} values.

The performance of the objective measures was evaluated on the $DMdnOS_p$ and $DMdnOS$ scores using the Pearson linear correlation coefficient (PLCC) to evaluate the prediction accuracy, the root-mean-square error (RMSE) to evaluate the absolute prediction error, and the Outlier Ratio (OR) to evaluate prediction consistency, as per ITU-T Recommendations,[15,16] and the Spearman rank order correlation coefficient (SROCC) to evaluate prediction monotonicity.[16] Also reported is the coefficient of determination (R^2), which measures the amount of variance explained by the model.

3. RESULTS

3.1 Subjective quality assessment

The quality rating of the laparoscopic sequences made by surgeons was inversely related to compression bitrate and was dependent on the scene, as shown in Figs. 2 and 3. Compression levels which surgeons judged significantly different in quality are shown in Fig. 3.1 as open triangles (Friedman post-hoc test with correction for multiple comparisons, p <0.05). Quality differences between the reference and compressed sequences were significantly different for the two largest compression level (2.8 and 1.8 Mbps) in scene A, with a 31% and 50% median decrease in perceived quality. In scene C, the largest compression level (1.8 Mbps) was significantly different from the reference, with a 34% median decrease in quality. There were no statistically significant differences between the reference and any of the compression levels in scenes B and D even though there was a median decrease in quality of 15% and 45% for scene B and 24% and 38% for scene D at the two largest compression levels.

In terms of scene, the surgeons rated scene C as having the best quality for all compression levels, A and B were rated lower but similarly to each other, and D was rated as having the worst quality for the three highest compression levels. At the largest compression levels, scenes A, B, and D were rated similarly.

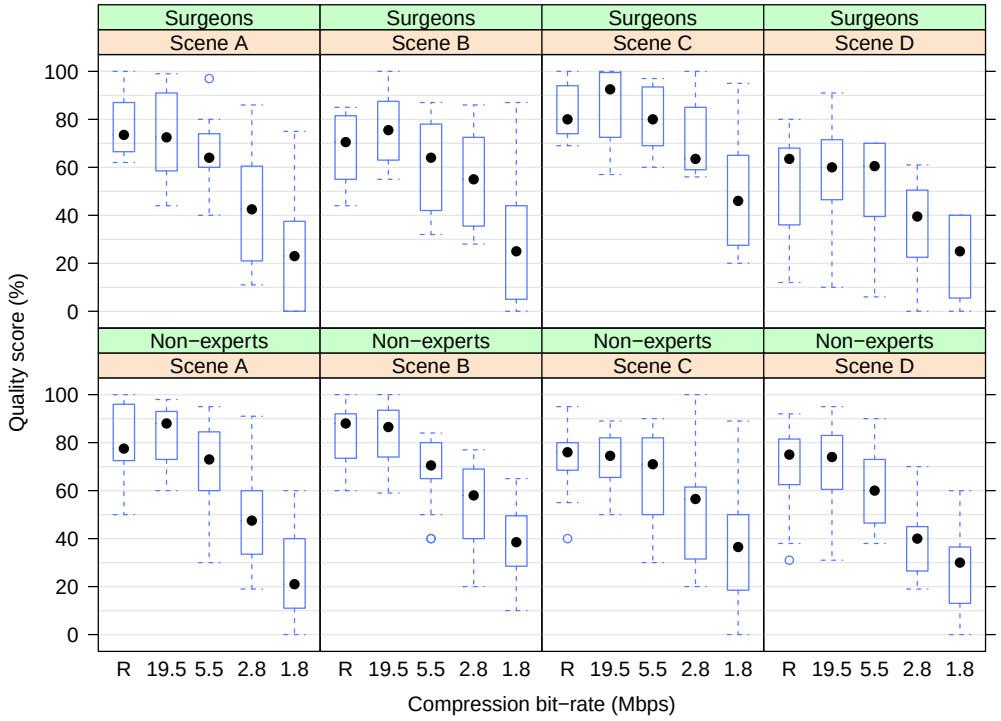


Figure 2. Quality scores from the SSCQE experiment represented as boxplots. Results for surgeons are on the top four graphs, non-experts on the bottom. The scores for a single scene are shown in each column. A score of 100% is "Excellent", 0% is "Poor". "R" is the reference (uncompressed) sequence.

Notably, the 19.5 Mbps compressed sequence was judged to be about 3% better in quality than the reference, averaged across all subjects for all scenes. Specifically, the median quality improvement was 5 and 12% in scenes B and C. In addition, there was a significant difference between the 19.5 and 1.8 Mbps compressed sequences for scenes B and D, with a 50% and 35% decrease in perceived quality at 1.8 Mbps.

The ratings of "suitability for surgery" indicated more significant differences than the quality rating. As illustrated in Fig. 3 with open triangles on the bottom graph, surgeons rated the 1.8 compressed sequence as significantly more unsuitable for use during surgery than the reference for all scenes (ranging from 30% to 48% reduction in quality), and the 2.8 Mbps as unsuitable for scene A (31% reduction). In addition, similar to the quality scores, the 19.5 Mbps compressed sequence was rated as being slightly more suitable for use during surgery than the reference in scenes A and C (3 and 2%). Spearman correlation between the surgeons' scores for quality and suitability for surgery was 0.97, and Kendall's tau was 0.88. The scatter plot is shown in Fig. 4a.

There were also important differences in quality ratings as a function of subject group. Non-experts consistently rated the two highest compression levels significantly different from the reference for all four scenes, with a decrease in perceived quality of approximately 20-35% at 2.8 Mbps and 35-50% at 1.8 Mbps. The compression levels significantly difference from the reference sequence, as judged by non-experts, are shown in Fig. 3.1 as open circles. The lack of significance in surgeons' scores at the highest compression levels was likely due to the slightly larger variance in their scores, seen on Fig. 2. Note also that for both groups, the variance in the scores increase with increasing compression level for scenes A-C. Non-experts rated the 19.5 Mbps 10% better than the reference for scenes A, and about 1% worse in the other three scenes, with a mean difference of 0% across all subjects and scenes.

Although the median scores of non-experts were higher than surgeons' scores for many of the sequences, between-subjects testing for the effect of expertise only showed significant differences for the reference sequence in scenes B and D (Wilcoxon rank sum test, $p < 0.05$), where non-experts rated the two reference sequences 17%

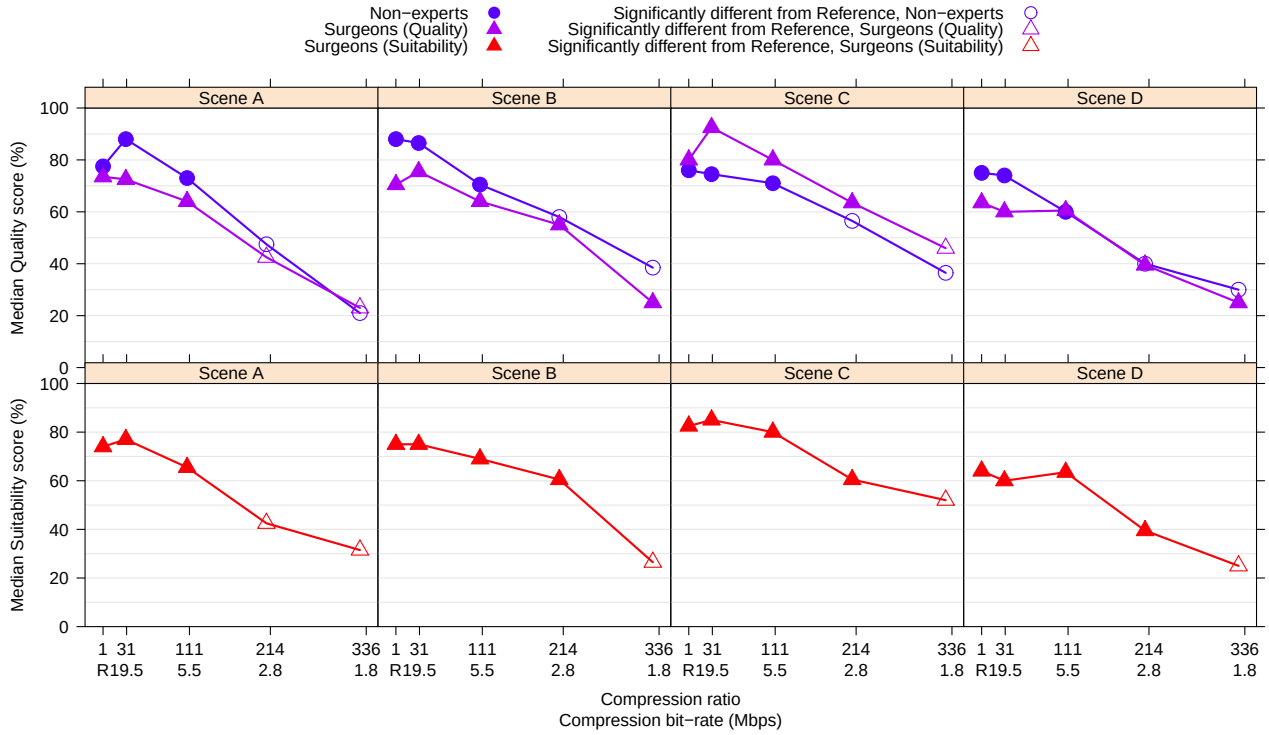
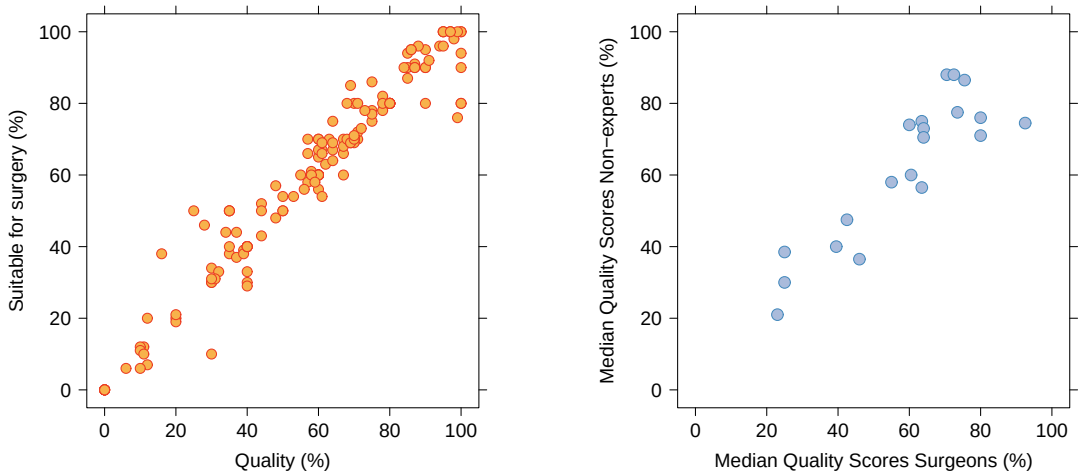


Figure 3. Median Quality (top graph) and "Suitability for surgery" scores (bottom graph) showing which compression levels were statistically significantly different (open symbols) from the reference sequence ("R"). Non-experts' scores are blue circles, surgeons' scores are purple triangles (Quality) and red triangles (Suitability). A score of 100% is "Excellent", 0% is "Poor". For example, in scene A, the perceived quality of the 2.8 and 1.8 Mbps sequences are significantly different from the reference for both subject groups.



(a) Surgeons: Quality vs. Suitability

(b) Median Quality: Surgeons vs. Non-experts

Figure 4. Scatter plots of (a) surgeons' scores, quality versus suitability for surgery: each point represents one subject per sequence, and (b) median quality scores for surgeons versus non-experts: each point represents one sequence.

Table 1. Performance of objective measures fit to each subject group

Subject group	Objective measure (Scenes included)		R^2	PLCC	SROCC	RMSE	OR (%)
Surgeons	VQM	(all)	0.94	0.97	0.94	0.27	6
	HDR-VDP-2	(all)	0.53	0.72	0.72	0.79	44
	HDR-VDP-2	(A,B)	0.98	0.99	1.00	0.22	0
	HDR-VDP-2	(C,D)	0.89	0.95	0.90	0.48	0
	PSNR	(all)	0.56	0.75	0.78	0.76	44
	PSNR	(A,B)	0.98	0.99	1.00	0.24	0
	PSNR	(C,D)	0.92	0.96	0.93	0.41	0
Non-experts	VQM	(all)	0.92	0.96	0.94	0.35	6
	HDR-VDP-2	(all)	0.60	0.77	0.77	0.77	44
	HDR-VDP-2	(A,B)	0.93	0.96	0.95	0.51	0
	HDR-VDP-2	(C,D)	0.99	1.00	0.95	0.12	0
	PSNR	(all)	0.64	0.80	0.83	0.73	44
	PSNR	(A,B)	0.93	0.96	0.95	0.53	0
	PSNR	(C,D)	0.99	0.99	0.98	0.15	0

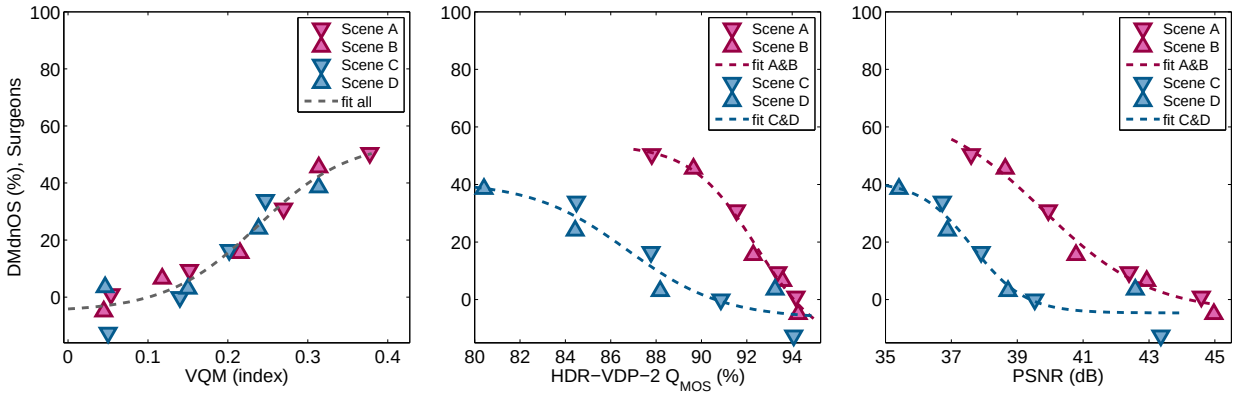
and 11% higher in quality. Similar to the surgeons, non-experts also rated scene D worse than the other reference scenes. Overall, the Spearman correlation between the median scores of expert and non-experts was 0.83, while Kendall's tau was 0.65; the scatter plot of the scores are shown in Fig. 4b. None of the subjective scores were detected as being outliers.

3.2 Objective quality assessment

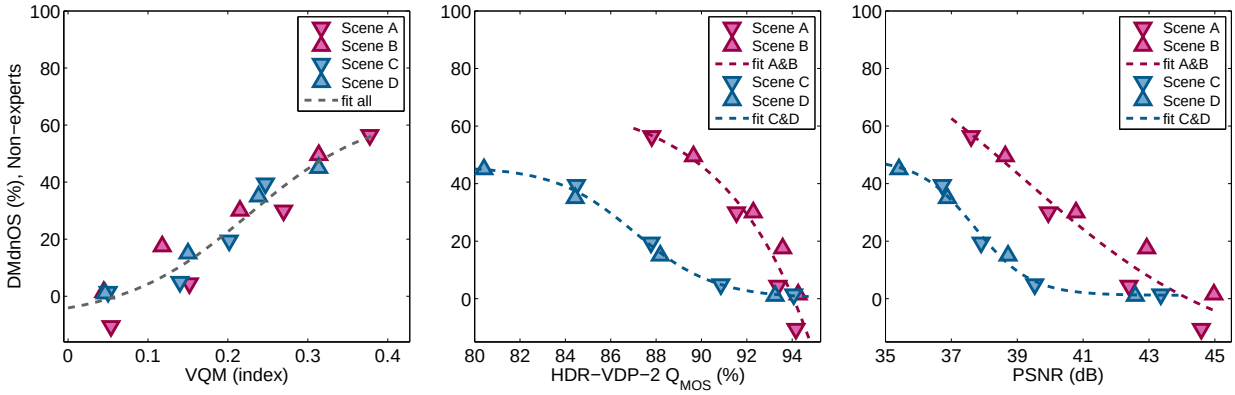
Of the three objective measures tested, only VQM showed a reasonable correlation with the subjective scores when all scenes were evaluated together, with R^2 and Pearson and Spearman correlations near 1, low RMSE, and outlier ratio near 0, as shown in Table 1. When the data were reanalyzed by separately fitting the scores by camera type (scenes A and B together, C and D together, as explained in section 2.1), then HDR-VDP-2 and PSNR also reached R^2 and correlations near 1, moderately low RMSE scores, and no outliers. The fitted curves with initial points are shown in Fig. 5. The performance of the measures for non-experts and surgeons is very similar for VQM. HDR-VDP-2 and PSNR perform slightly better for surgeons in scenes A&B and for non-experts in scenes C&D. The shape of the fitted curves is very similar for the two subject groups, but different for the two camera types. Despite the high correlations between objective and subjective scores, Fig. 5 illustrates that the objective measures are at times unable to differentiate the quality level between sequences that humans do. For example, VQM (Fig. 5, left plots) and HDR-VDP-2 (Fig. 5, middle plots) do not distinguish between the four scenes at the highest high quality level (DMdnOS near zero), whereas humans do. Also notable in Fig. 5 is that some of the curves dip below zero on the y-axis. Since DMdnOS is the difference in quality scores between the reference and compressed sequence, and subjects rated the 19.5 Mbps sequence better than the reference for some scenes, this is expected behavior.

4. DISCUSSION

This study's findings suggest that full HD resolution laparoscopic video sequences may be compressed up to 19.5 Mbps (using the H.264/AVC IPPP-CBR codec) with no visible loss in perceived quality, and as low as 5.5 Mbps may be suitable for use during surgical procedures with no or limited loss in perceived quality. This translates to savings of approximately 30 to 100 times in storage space and bandwidth. These findings are in line with the



(a) Objective scores fit to surgeons' scores



(b) Objective scores fit to non-experts' scores

Figure 5. Plots of objective measure scores (x-axis) versus difference median opinion scores (DMdnOS, y-axis) for the three objective measures. Each point represents one sequence; the dashed line indicates the best-fit non-linear regression curve. VQM was regressed over all scenes; HDR-VDP-2 and PSNR by camera type. Goodness-of-fit and performance values are given in Table 1.

study[3] indicating that MPEG-2 [4:2:2] 40 Mbps compressed sequences were indistinguishable from the reference video, and that as low as MPEG-2 [4:2:0] 8 Mbps sequences were potentially acceptable for diagnostic use.

The results of this study also suggest that quality preferences are dependent on the type of subject: while non-experts judged the effect of compression on quality very similarly across the four scenes, surgeons were sensitive to content. Surgeons likely took into account the ability to appreciate specific anatomical structures when assessing the quality. For example, surgeons rated the reference sequence of scene D much lower in quality than other reference scenes, possibly due to the presence of smoke and bleeding in the image, which partially prevented the visualization of the underlying anatomy. Conversely, non-experts possibly did not appreciate the loss of anatomical detail and instead concentrated on compression artifacts visible in the smoke. In addition, image quality surgeons consistently indicated that the 2.8 Mbps sequence was of significantly poorer quality than the reference; while surgeons also judged those sequences as much worse than the reference on average, significance testing did not reveal differences for some scenes. One possible reason is the reduced statistical power of non-parametric tests, compared to parametric tests. Alternatively, lack of significance may have be due to the large inter-subject variances in the surgeons' quality scores, which in turn could have been caused by insufficient training in the task or the use of the scoring scale, varying experience levels, or lack of a consensus on the quality criteria. While the Spearman correlation between the two subject groups was high, Kendall's tau of 0.65 indicates that the probability of agreement between the groups is only moderate. Without further study into the reasons for these differences, we suggest that non-experts should not be used a surrogates for surgeons' quality judgments.

It was surprising that the 19.5 Mbps scene was rated better than the reference in some scenes. One explanation for the improved perceptual quality of the compressed sequence is that compression at high bit-rates removes high-frequency noise while preserving important spatial details; this same effect has been noted in a study on compressed chest radiographs.[17] Another explanation is the inherent variability in the quality judgments, as the higher rating was not dependent on subject or scene. Therefore, approaches that limit both intra- and inter-subject variability may improve the reliability of experimental findings: double stimulus, paired comparison, or forced-choice protocols, which may be more time-consuming but allow the subject to anchor their decisions against a reference image; and additional training in the task and use of the scale.

This study also found minimal effect of task: surgeons' judgments on the overall quality and "suitability" of the sequences for use during surgery were highly correlated and concordant. Although the "suitability" task may seem more relevant to the clinical application, both tasks were likely interpreted in the same context: a high quality sequence will be more suitable for use during surgery. Future studies that attempt to elicit additional information about the suitability of the task for surgery could consider alternative testing scenarios - for example by asking for feedback about the suitability of the quality while conducting a (real or simulated) surgical procedure. However, these tasks are all subjective: they measure the subject's *opinion* of the *appearance* of the sequence using some criteria. Classification tasks such as detection and discrimination, on the other hand, measure the effect of the degradation on the *performance* of the task, and are the preferred method of quality assessment in medical imaging*. A follow-up study to this one might, for example, assess the detection performance using a specific anatomical structure (e.g. vessel or lesion) in compressed laparoscopic video sequences via receiver operating characteristic (ROC) methodology. Since ROC studies require multiple stimuli per condition per subject - more than the four scenes reported here - the type of experiment reported in this study could potentially be used to preselect a set of compression bit-rates of interest to build the ROC experiment. Then, a detection task could be used to assess the clinical implications of compression levels at which subjective quality begins to degrade (2.8 Mbps to 5.5 Mbps in this study). Currently there is little investigation into the concordance of appearance and performance based measures of medical image quality.

Finally, while objective quality measures seemed to predict the quality judgments of surgeons, they did not always differentiate between sequences that human subjects did. Furthermore, two measures (HDR-VDP-2 and PSNR) gave very different results depending on the camera type: the measures rated the compressed sequences of scenes A and B higher than those of C and D although the subjects rated them similarly. One possible explanation might be that the chip-on-tip camera used for scenes C and D generates images that have more noise than the standard camera type – compression removes the noise, which results in larger measured differences between the reference and compressed sequences, causing lower quality scores for this camera type and completely different quality curves for the two camera types. In addition, the developers of HDR-VDP-2 report that the quality measure's visible difference predictor was calibrated to the LIVE image database,[11] which may have caused biases towards the types of images in that database. In addition, the authors themselves report that the quality measure is "good at ranking each distortion type individually, but is less successful at differentiating the quality between the distortions"[11] - the differences between scene types may also have been affected by the same phenomenon. The VQM metric does not appear to suffer from this problem. Future studies should take into consideration that camera types, and potentially even different anatomical regions or clinical tasks, may not be evaluated equivalently by objective measures.

This study had several additional limitations. A total of four scenes - two per camera type - were evaluated. One scene was inherently degraded due to the presence of smoke and blood. The small number of scenes and lack of diversity in content limit the extent to which the results can be generalized to all laparoscopic procedures. In addition, only one type of lossy compression codec was evaluated (H.264/AVC optimized for low latency). Different compression parameters will generate sequences with different perceived quality even at the same compression bit-rate, due to selectable codec parameters such as the profile, level, motion estimation search pattern (me) and range (merange), and the use of B frames. Therefore, it can be challenging to compare results

*Typically, classification tasks - even when conducted by humans - are referred to as *objective* tasks in medical image quality assessment. However, to avoid confusion, in this paper we use the term *objective* in the general video quality assessment sense: metrics which assess the *appearance* or *fidelity* of the image, and not the performance on a particular clinical task.

of QA studies if the compression settings are not known. The type of codec (e.g. MPEG-2, MPEG-4 Part 2, H.264/AVC, or HEVC) will also greatly influence the perceived quality of sequences compressed at the same bit-rate, and cannot be directly compared at a particular compression bit-rate or ratio; this problem has not yet been solved for still-image compression of medical images.[18] Finally, objective video quality measures could ideally be used to benchmark the perceived quality of different compression codecs and bit-rates, but as this study suggests, the algorithms may be sensitive to image and video artifacts that humans do not perceive, and vice-versa, rendering them currently unsuitable for such a task.

5. CONCLUSION

This study, conducted under a general video quality assessment paradigm, suggests that laparoscopic video sequences may be lossy compressed approximately 30 to 100 times using H.264/AVC optimized for low latency, without sacrificing perceived visual quality. This codec shows promise as a suitable candidate for the storage and transmission of laparoscopic video at compression bit-rates of 19.5 to 5.5 Mbps, potentially enabling real-time streaming of surgical procedures even over wireless networks.

The high correlation between surgeons' scores for "suitability for use during surgery" and quality suggest that the two tasks were likely interchangeable in this experimental context. The findings also suggest a strong effect of subject expertise: surgeons appeared to be sensitive to content but had large variances in quality scores, whereas non-experts judged all scenes similarly and over-estimated the quality of some sequences. Therefore, non-experts should not be used to estimate quality preferences of surgeons. The three objective measures had moderate to high correlation with subjective scores, especially when analyzed separately by camera type. However, the measures had difficulty in distinguishing the quality between sequences at the same compression level but from different scenes, indicating that these quality measures are likely sensitive to compression artifacts or features in the original sequences which human subjects are not. Without further investigation, this finding limits the utility of current objective measures for comparing the quality of different compression parameters and codecs in laparoscopic videos. In addition, we recommend that authors report the compression parameters used in QA studies to facilitate comparison between different compression parameters and codecs.

Several limitations of this study should be addressed in the future. The assessment of additional scenes for different laparoscopic camera types would strengthen the findings of this study and help determine the shortcomings of the objective measures. Testing of additional compression H.264/AVC parameters or other codecs would also better assist the design of objective measures intended for medical video applications. Techniques to reduce intra- and inter-subject variability in quality scores would increase the power of the study, for example via an alternative QA protocol, additional training in the use of the scoring scale, or ensuring consensus on the quality criteria.

Finally, these results should be considered in the context of this experiment: surgeons were asked for their *subjective opinion* on the quality and "suitability for use during surgery" of the sequences. While 19.5 to 5.5 Mbps compression may be suitable for remote viewing or storage purposes, the effect of lossy compression on diagnostic and clinical outcomes when used during surgical procedures should be evaluated with a performance or task-based quality assessment paradigm. Future work may, for example, assess the surgeon's ability to detect specific anatomical or pathological features critical to laparoscopic surgery when the procedure is conducted at different compression bit-rates.

6. ACKNOWLEDGMENTS

We would like to thank all the participants who volunteered their time to participate in this experiment.

This work was financially supported by the Telesurgery project, a project co-funded by iMinds, a research institute founded by the Flemish Government. Companies and organizations involved in the project are BARCO, Unilabs Teleradiology BVBA, and CandiT-Media BVBA, with project support of IWT.

REFERENCES

[1] Juurlink, B., Alvarez-Mesa, M., Chi, C., Azevedo, A., Meenderinck, C., and Ramirez, A., "Understanding the application: An overview of the H.264 standard," in [*Scalable Parallel Programming Applied to H.264/AVC Decoding*], *SpringerBriefs in Computer Science*, 5–15, Springer New York (2012).

[2] Nouri, N., Abraham, D., Moureaux, J.-M., Dufaut, M., Hubert, J., and Perez, M., "Subjective MPEG2 compressed video quality assessment: Application to tele-surgery," in [*Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*], 764 –767 (2010).

[3] Rabenstein, T., Maiss, J., Naegele-Jackson, S., Liebl, K., Hengstenberg, T., Radespiel-Trger, M., Holleczek, P., Hahn, E. G., and Sackmann, M., "Tele-endoscopy: Influence of data compression, bandwidth and simulated impairments on the usability of real-time digital video endoscopy transmissions for medical diagnoses," *Endoscopy* **34**, 703–710 (2002).

[4] Przelaskowski, A. and Jozwiak, R., "Compression of bronchoscopy video: Coding usefulness and efficiency assessment," in [*Information Technologies in Biomedicine*], Pietka, E. and Kawa, J., eds., *Advances in Soft Computing* **47**, 208–216, Springer Berlin Heidelberg (2008).

[5] ITU-R Recommendation BT.709-5, "Parameter values for the hdtv standards for production and international programme exchange." International Telecommunications Union (April 2002).

[6] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," (May 2003).

[7] VideoLan Organization, *x264 core:119 r2106 07efeb4*. http://www.videolan.org/x264.html.

[8] Balakhnin, A., *YADIF, version 1.7*. http://avisynth.org.ru/yadif/yadif.html.

[9] ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures." International Telecommunications Union (January 2012).

[10] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting* **50**, 312–322 (September 2004).

[11] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W., "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," in [*ACM Transactions on Graphics (Proc. of SIGGRAPH'11)*], **30**(4), 40 (2011).

[12] Wolf, S. and Pinson, M., "Video quality measurement techniques," Tech. Rep. NTIA Report TR-02-392, Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, 325 Broadway, Boulder, Colorado, 80305, USA (June 2002).

[13] Galili, T., "friedman.test.with.post.hoc, performing the post-hoc tests of: Wilcoxon-Nemenyi-McDonald-Thompson test, Hollander & Wolfe (1999), page 295," (http://www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code/, Last accessed: August 06, 2013).

[14] R Core Team, *R: A Language and Environment for Statistical Computing, v3.0.2*. R Foundation for Statistical Computing, Vienna, Austria (2013).

[15] ITU-T Recommendation P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," (July 2012).

[16] "Final report from the video quality experts group on the validation of objective models of video quality assessment," tech. rep. (2000).

[17] Slone, R. M., Foos, D. H., Whiting, B. R., Muka, E., Rubin, D. A., Pilgram, T. K., Kohm, K. S., Young, S. S., Ho, P., and Hendrickson, D. D., "Assessment of visually lossless irreversible image compression: Comparison of three methods by using an image-comparison workstation," *Radiology* **215**(2), 543–553 (2000). PMID: 10796938.

[18] Fidler, A., Skaleric, U., and Likar, B., "The impact of image information on compressibility and degradation in medical image compression," *Medical Physics* **33**(8), 2832–2838 (2006).