Investigation of the Impact of Compression on the Perceptual Quality of Laparoscopic Videos

Bernd Münzer∗, Klaus Schoeffmann∗, Laszlo Bőszörményi∗, JF Smulders† and Jack J. Jakimowicz†‡
∗Institute of Information Technology / Lakeside Labs, Klagenfurt University, Austria
†Department of Surgery, Catharina Hospital Eindhoven, The Netherlands
‡Faculty of Industrial Design Engineering, Delft University of Technology, The Netherlands
{bernd, ks, laszlo}@itec.aau.at, frans.smulders@cze.nl, jakimowi@planet.nl

Abstract—In recent years it has become common practice to archive video recordings of laparoscopic surgeries for documentation purposes and for retrospective review. Typically, the videos are captured in High Definition (HD) format but encoded with legacy coding standards like MPEG-2 requiring an enormous storage capacity. In this paper we present the results of a subjective quality assessment study with 37 medical experts. We identify appropriate encoding configurations for the H.264/AVC coding standard to guarantee a visually lossless quality with a significant bitrate reduction. Further, we show that it is not necessary to capture the highest possible quality for documentation and retrospective analysis. A lower technical quality with a substantially lower bitrate still provides sufficient semantic quality. We finally present basic recommendations for an efficient encoding strategy with an appropriate tradeoff between visual quality and bitrate.

I. INTRODUCTION

Minimal invasive surgery, also referred to as “keyhole” surgery or in the case of abdominal operations laparoscopic surgery, has gained a more and more important role in the armamentarium of a surgeon (e.g., removal of the gallblader). As opposed to open surgery, only a few small incisions are necessary to insert the required surgical instruments. The surgical field is viewed indirectly on a monitor that displays a video signal captured with a digital camera mounted on a so-called laparoscope (a special type of endoscope). In recent years, it has become common to record such video streams for documentation purposes, education, research and retrospective analysis, e.g., to assess the skills of surgeons. In some countries, the video documentation is even mandatory already and it is foreseeable that in the near future most interventions will be recorded.

A comprehensive video documentation produces large video archives that are hard to manage. One of the main problems is the huge data volume and the associated storage space requirements. In recent years, there has been a trend towards High Definition (HD) video which is currently establishing as common practice both for the real-time display during the surgery and for video recording. Commercially available video documentation systems often use the MPEG-2 video format in constant bitrate mode (CBR) although it has not been designed to handle HD content. With our reference system, recording in HD requires a bitrate of 20 Mb/s. This means that one hour of video consumes about 9 GB. In a typical hospital about 20-25 laparoscopic procedures with an average duration between one and two hours are carried out per day. In this example, comprehensive video recording would lead to a data volume of about 300 GB per day, 2 TB per week and 100 TB per year.

Hospital operators are not able to provide the necessary infrastructure for storing and accessing such huge amounts of data. Hence, videos are often archived on external storage media like DVDs and can only be retrieved and used to a very restricted extent. The long-term goal of our research is to build an efficient content-based management and retrieval system for endoscopic video archives that includes advanced features like summarization [1], detection of irrelevant segments [2], semantic segmentation or similarity search and exploits the huge potential benefit of this big data. A very important precondition for such a system is an adequate video encoding strategy with a fair tradeoff between video quality and storage space requirements that enables a balanced cost-benefit ratio.

The first step to reduce the storage requirements is to migrate to a more recent coding standard. We propose to use H.264/AVC [3] because it is the current state-of-the-art standard for HD video. It is a successor of MPEG-2 and generally doubles the coding efficiency. The second and more challenging step is to find an appropriate tradeoff between video quality and bitrate. Surgeons generally agree that while high fidelity HD quality is very expedient during the surgical procedure, it is not necessary to store each and every video in the original high quality for everyday documentation purposes (unless they are intended for special use cases like education, training or research). This is especially true for interventions which are mainly therapeutic and do not have a major diagnostic component. Unfortunately, it is not exactly known what minimum quality level is required for a laparoscopic...
video to still retain all medically relevant information. This question has not been addressed in the literature so far and no guidelines or recommendations are publicly available.

To approach this problem, we have to differentiate between different perspectives on video quality. The most common distinction is made between subjective and objective quality. The subjective quality relates to how a video is perceived by the Human Visual System (HVS) of an observer and can only be measured by laborious and expensive user studies. The objective quality on the other hand is measured by mathematical models that try to approximate the subjective quality at much lower cost. A lot of research has been conducted in this field [4], but objective models are not able to reliably predict the subjective quality perception, especially if different coding standards and resolutions are involved. Hence, we can not rely on them in such a special and critical domain as laparoscopy.

In this paper, we additionally consider three other aspects of video quality. The content quality relates to the amount of information that the video contains, irrespective of the technical quality. As it is determined at acquisition time, we can not influence the content quality. The technical quality is specified by encoding parameters and determines the required bitrate to encode the video. The storage requirement of a video is directly affected by the bitrate (size = bitrate × duration). The primary parameters are resolution, quantization and frame rate. The semantic quality [5], sometimes also called utility, describes to what extent the relevant information of the video is conveyed to the observer. A more compressed version of a video sequence with a lower technical as well as subjective quality can still have the same semantic quality as long as all relevant information can be recognized. The semantic quality of laparoscopic videos can only be assessed by domain experts (surgeons). For non-experts it is impossible to judge which image details contain medically relevant information.

In this work, we pose the following research questions:

1) To what extent can laparoscopic videos be compressed without a subjectively noticeable quality loss?
2) Is it further feasible to reach a significantly higher compression rate without essential loss of semantic quality?

We conducted a so-called subjective quality assessment study with 37 surgeons and surgical residents. To the best of our knowledge, this is the first study investigating the correlation of technical and semantic video quality in this field. It consists of two sessions with different test conditions and assessment methods, each tailored to one research question. Additionally, we investigate the impact of image downscaling versus coarser quantization and the impact of a surgeons’ experience on their quality requirements. Finally, we give encoding recommendations for different scenarios based on our findings and point out future work directions.

II. SUBJECTIVE QUALITY ASSESSMENT

This section describes the details of the user study setup, selection of video sequences and participants as well as test methods employed. Most aspects are largely based on the well-established recommendations for subjective video quality assessment of the ITU (International Telecommunication Union) [6] [7] with minor modifications.

A. Participants

In order to obtain statistically robust results, we conducted the study with 37 participants (7 female, 30 male). All of them are medical domain experts and are no experienced assessors or multimedia experts. 18 of them are surgical residents (surgeons who have received their medical degree but are still in training) and 19 are experienced surgeons. The average age of the participants is 40 with a standard deviation of 12.8. Each participant attended two test sessions of about 30 minutes each on different days (with the exception of one single participant who was only able to participate in the first session due to time constraints). Prior to the study, the participants received a written summary of the type of assessment, the opinion scale and the presentation of the test sequences as an introduction into the goals and procedure of the study.

B. Video Sequences

We use in total 48 distinct video sequences (VS) with a length of 10 seconds each. They show representative and relevant scenes of laparoscopic procedures which were carefully selected by medical domain experts. Special care was taken to select a diversified dataset with clips from different procedures and operation phases (e.g., overview scenes, cutting scenes, suturing scenes, high motion scenes etc.). Example frames are depicted in Figure 1. The resolution of the original sequences, henceforth referred to as reference sequence (RS), is 1920x1080 (Full HD). The temporal resolution is 25 fps (frames per second). It was not possible to obtain the perfectly original (uncompressed) video signal for technical reasons, but the reference sequences were recorded at a very low compression rate that can be considered to be visually lossless (MPEG-2 in constant bitrate mode with a bitrate of 35 Mb/s).

For each RS a number of test sequences (TS) have been created using the H.264/AVC coding standard. Each TS represents one test condition (TC). Test conditions are characterized by resolution and quantization coarseness. A coarse quantization corresponds to a high compression rate while a fine quantization only leads to low compression. The TCS are different for the two study sessions and are described in detail in section II-D. In each session, each VS is presented exactly once, but the association of TC to VS is different for each participant. This is very important because it minimizes the impact of individual videos and their underlying content quality on the ratings (also known as scene bias).

C. Test Setup and Assessment Method

The technical setup is the same for both sessions: The participant sits in front of a HD wide screen monitor with a screen diagonal of 27 inches and a resolution of 1920x1080. The monitors we use have been approved as medical devices. The ambient light in the room is turned off. A notebook is positioned right beside the monitor and shows a discrete quality scale from 0 to 100. It is uniformly subdivided into five sub-ranges: Excellent, Good, Satisfactory, Sufficient, Insufficient. The semantics of the rating scale are additionally visualized by a color gradient. The categories are similar to the ones proposed in [6] but had to be adapted to our assessment objectives. The technical setup and the rating screens for the two sessions are illustrated in Figure 2.
A number of video sequences are presented in full screen mode. Test videos with a reduced resolution are scaled to full screen using bilinear interpolation which is a common interpolation method in standard video players. We did not use more advanced methods (e.g., bicubic or lanczos) on purpose because they might not be available in a practical scenario. The participant is asked to carefully watch the presentations and then rate their perceived subjective video quality by moving a slider with the mouse. They are explicitly instructed to do so with respect to the purpose of retrospective review, not for the real-time display during the surgery. The participants decide for themselves when to continue with the next task and can take a short rest after any task, if necessary. In the middle of the test, a short break of 30 seconds is enforced.

1) Session 1: The objective of the first session is to find appropriate encoding parameters for maximum compression without any subjectively noticeable quality degradation. For that purpose, we employ the double-stimulus continuous quality-scale method (DSCQS), variant I, with minor modifications. It has been recommended and standardized by the ITU [6] and is one of the most used assessment methods. It is especially suited for the detection of small differences due to its direct coupling of reference and test sequence. In each task a video sequence pair (reference and test) is presented. At each point in time exactly one of the two sequences is visible. The participant can switch between the two sequences at any time by pressing a button. They are not told which of the two is the reference version, the order is determined randomly. After a short break of 2 seconds, the presentation is repeated once to give the participants the chance to stabilize their opinion. In the rating step, the participant rates the perceived quality of both versions.

2) Session 2: The second session is intended to investigate how a reduction of the technical quality affects the semantic quality. For that purpose, we use a single-stimulus quality assessment method with hidden reference, similar to the ACR-HR method [7], but with a continuous rating scale as in session 1. Each task consists of two subtasks that are presented independently. They are not shown consecutively but at least three other subtasks are shown in between. One subtask only shows the TS while the other only shows the RS. The sequences are not repeated. The participant does not know that there is a reference for each test sequence but rates the sequence of every subtask separately as if it were a TS.

3) Training: In both sessions, three training presentations are shown at the beginning of the session. They are intended to make the participant familiar with the task, the user interface and the range of qualities to be expected. These sequences are the same for all participants. They are not taken into account for the evaluation.

4) Reliability Test: Additionally, a small number (4) of duplicate pair tasks are used at random positions in the session to validate the intra-subject reliability of a participant. They contain two identical sequences (either reference or test). If the participant is reliable, it is likely that both ratings are very similar. If they are very different it might be an indication that the participant rates in a rather random fashion.

D. Test conditions

In each session, we evaluate 12 different test conditions with 4 replications each, resulting in 48 tasks for each participant. The replications are important to measure the intra-subject reliability. For each task, a different VS is used. The association between VS and TC is randomized for every participant, but with the constraint that the TCs are evenly distributed over all VS. The presentation order of the VS is also randomized to minimize the impact of context. Some additional constraints are imposed on the presentation order, to ensure a balanced distribution of the TCs over the session and further reduce the context bias. The employed test conditions are stated in Table I. Their average bitrates are illustrated in Figure 3.

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1920x1080</td>
<td>20, 22, 24, 26, 28</td>
<td>24, 26, 28</td>
</tr>
<tr>
<td>1280x720</td>
<td>18, 20, 22, 24, 26</td>
<td>20, 22, 24, 26</td>
</tr>
<tr>
<td>960x540</td>
<td>18</td>
<td>18, 21, 24</td>
</tr>
<tr>
<td>640x360</td>
<td>18</td>
<td>18, 21</td>
</tr>
</tbody>
</table>
The parameters for the TCs have been selected empirically. We use 4 different resolutions: the two common HD resolutions (1920x1080 and 1280x720) and for each a lower resolution with half the number of pixels in both dimensions (960x540 and 640x360). For each resolution we use a number of crf (constant rate factor) values. Crf is the most common rate control mechanism of the x264 encoder and is designed for a constant quality and variable bitrate. The crf range is 0-51 with a higher crf value meaning coarser quantization and thus lower quality. However, only the range of 18-28 is considered “subjectively sane” with crf 18 being regarded as visually lossless. We tried to achieve a balanced coverage of the bitrate spectrum and a bitrate interlacement for different resolutions. That means that we also have TCs with lower resolution but higher or comparable bitrate. We conducted an internal pretest with 5 multimedia experts without medical expertise to obtain preliminary indicative results according to which we slightly refined the test conditions.

III. RESULTS

A. Pre-processing

In user studies it is common to pre-process the raw data, especially in terms of outlier removal. We applied the outlier detection method proposed in [6], but no outliers were found. We experimented with several other approaches for outlier detection (e.g., by analyzing the duplicate pair tasks) but finally decided to use all data because the effect of outlier removal was very small. This is a good indication that the chosen number of participants is high enough to innately tolerate outliers.

According to [7], all reference sequences should have “good” or “excellent” quality. Figure 4 depicts the average ratings of all reference sequences for both sessions. For six reference sequences, the average rating is below 60 in both sessions which only corresponds to a “satisfactory” quality. Reference sequences with low quality distort the results because the content quality negatively affects the rating. We removed all ratings for these sequences to obtain a more stable result. Note that this does not affect the uniform distribution of test conditions. In this context, we made the interesting observation that the reference ratings in session 1 are considerably higher than in session 2. We assume that this is due to the different assessment methods. In session 1, participants know that one of the two versions has to be the reference while in session 2, they do not know that reference videos exist at all.

Fig. 4. Average rating of the reference sequences for session 1 and 2.

B. Research question 1

To investigate our first research question (to what extent can laparoscopic videos be compressed without a subjectively noticeable quality loss?) we are interested in the rating difference between reference and test sequence rather than the absolute rating. If the objective quality loss introduced by the compression is subjectively unnoticeable, this value should be around zero.

For every completed task of session 1, we calculate the rating difference between reference and test sequence. We aggregate these values to obtain the Difference Mean Opinion Score (DMOS) for each test condition. Additionally, we calculate the 95% confidence intervals. We do not use normalization techniques (e.g. z-scores) because the meaning of the values would get lost. Figure 5 illustrates the obtained data and contrasts it with the average bitrate of the test conditions.

For the first four test conditions (1920x1080 with crf 20-26) we can observe a negative DMOS (below the red line). This means that the test video was rated even better than the reference video. Although the bitrates of these four TCs range from 23 Mb/s to 8 Mb/s no significant quality difference can be noticed. In full resolution, only the coarsest quantization (crf 28) has a positive DMOS (2.1), although still quite small.

The next smaller resolution (1280x720) shows DMOS scores between 5 and 8.6. This means that the quality degradation is clearly noticeable with this resolution, even with very low compression (fine quantization). Note that the test condition with the lowest compression in this resolution consumes a significantly higher bitrate than full resolution with maximum compression, but still has a larger quality loss. We can again observe that the DMOS for fine to medium quantization is quite constant despite large bitrate differences. Only very coarse quantization leads to a higher DMOS.

For the two lower resolutions (960x540 and 640x360), we observe a significant increase of the DMOS (16.3 and 30.4), meaning that the quality degradation is obvious to the medical domain experts.

Fig. 5. Results of session 1. The orange bars visualize the average bitrate of the 12 test conditions. The blue items depict the DMOS with 95% confidence intervals. Note that two different scales and consequently two y-axes are used. The red line highlights the zero rating difference.

1https://trac.ffmpeg.org/wiki/x264EncodingGuide
We can conclude that it is indeed possible to compress laparoscopic HD videos to a high extent without compromising the subjectively perceived quality. By using the H.264/AVC coding standard, the perceived quality is even slightly enhanced compared to the reference version. We assume that this is on the one hand due to coding optimizations and advanced features of H.264/AVC and on the other hand due to a slight reduction of the noise level caused by the quantization.

One important finding of our study is that it is necessary to retain the Full HD resolution. If the resolution is reduced, medical domain experts can definitely identify the degradation. Even a very coarse quantization in Full HD is rated better than the next lower resolution with very fine quantization, although the bitrate is considerably higher in the latter case.

C. Research question 2

To investigate our second research question (is it further feasible to reach a significantly higher compression rate without essential loss of semantic quality?) we use the ratings for the test sequences of session 2. We calculate the Mean Opinion Score (MOS) for each test configuration by averaging the individual ratings. We also compute the 95% confidence intervals. Additionally, we compute the average rating for the reference sequences as a basis for comparison. It is important to note that the participants did not know that reference sequences were included in session 2. They rated each task as if it were a test sequence. Figure 6 shows the results and again contrasts them with the average bitrate of the test conditions.

We can observe that all used test conditions achieve good results. Even the worst test condition with the lowest bitrate (640x360 with crf 21, 1.4 Mb/s) achieves a MOS of 41.3 which is at the boundary between sufficient and satisfactory on our rating scheme. This confirms our expectation that a low technical quality still provides an acceptable semantic quality and would easily be sufficient for archiving. Compared to our reference system that uses MPEG-2 with 20 Mbit/s this would be a bitrate improvement by a factor of 14.

The test conditions with full resolution confirm our findings from session 1. Even the most compressed (most coarsely quantized: crf 28) test condition with full resolution obtains a MOS that is slightly higher than the average reference rating. In session 1, the participants were able to detect the quality loss of this test condition because of the direct coupling of reference and test sequence. However, in session 2 this was not the case because they were presented independently.

Furthermore, we can see that the MOS within resolutions is relatively stable despite large bitrate differences. On the other side, a reduction of the resolution always leads to a lower MOS, even if the bitrate is significantly higher. As an example, the resolution 1280x720 with crf 26 has a MOS of 59.5 and a bitrate of 2.5 Mb/s while the resolution 960x540 with crf 18 only has a MOS of 57.7 but a bitrate of 8.5 Mb/s. If we compare test configurations with different resolutions but similar bitrates we can see that the higher resolution always has a higher MOS (e.g., 960x540 with crf 24 and 640x360 with crf 21). In summary, we can state that downscaling entails a higher quality loss than quantization. In other words, it is better to apply coarser quantization before reducing the resolution.

D. Impact of Experience

We further compare the ratings of the surgical residents and the experienced surgeons to find out if their perception of quality differs. For this comparison we use all test conditions of both sessions. The results are illustrated in Figure 7. We can see that the MOS of the residents is considerably lower for almost all test configurations. The average difference is 3.4 points. Also the average reference rating is 3.5 points lower. This shows that experienced surgeons seem to have lower quality requirements. The reason could be that they can interpolate certain details because of their experience. Another reason could be that residents are spoiled by high quality videos while experienced surgeons already had to cope with lower quality equipment in the past and therefore have a different appreciation for video quality.
IV. Encoding Recommendations

In the following, we summarize the findings of our study to provide basic recommendations for an efficient encoding of HD videos for a laparoscopic video archive with H.264/AVC. Our recommended encoding configurations for the x264 encoder are summarized in Table II. The storage space ratio is calculated with respect to the current practice of our reference system. Please note that our study is fragmentary at the low end of the bitrate spectrum as it rather focuses on the high end as a first step. Further studies should be conducted to complement our findings and in particular investigate the boundary to insufficient semantic quality.

<table>
<thead>
<tr>
<th>Quality level</th>
<th>Resolution</th>
<th>Crf</th>
<th>Bitrate</th>
<th>Storage requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current practice</td>
<td>1920x1080</td>
<td>MPEG-2</td>
<td>20 Mb/s</td>
<td>100%</td>
</tr>
<tr>
<td>Visually lossless</td>
<td>1920x1080</td>
<td>26</td>
<td>8 Mb/s</td>
<td>40%</td>
</tr>
<tr>
<td>Good quality</td>
<td>1280x720</td>
<td>26</td>
<td>2.5 Mb/s</td>
<td>12.5%</td>
</tr>
<tr>
<td>Acceptable quality</td>
<td>640x360</td>
<td>21</td>
<td>1.4 Mb/s</td>
<td>7.14%</td>
</tr>
</tbody>
</table>

A. Scenario 1

If the given quality requirements demand that the original video signal must be encoded without quality loss (e.g., for legal reasons or because they are intended for special use cases like education or research), it is essential to retain the full resolution. However, our study provides evidence that even strong quantization does not reduce the subjective quality. Usage of the H.264/AVC coding standard even slightly increases the perceived quality. For the x264 encoder, the crf value of 26 is an appropriate choice to encode laparoscopic HD videos without perceivable quality degradation. As a rule of thumb we can state that a bitrate of about 8 Mbit/s is sufficient for that purpose. Encoding with a higher bitrate does not provide any gain in subjective quality. Compared to the current practice of many commercial systems (MPEG-2 with 20 Mbit/s) this is a bitrate improvement by a factor of 2.5.

B. Scenario 2

If the quality requirements are less restrictive and only demand that the semantic quality of the videos must be guaranteed, the data volume can be reduced to a much larger extent. The challenge is to find a fair tradeoff between video quality and bitrate. If the full HD resolution is retained, it is possible to apply even more quantization with only minor subjective quality degradation. With a crf of 28 the bitrate can be reduced to 5 Mb/s (a reduction factor of 4). In a direct comparison to the original, the degradation is just noticeable, but the semantic quality is still very high.

A more significant bitrate reduction (2.5 Mb/s, reduction by a factor of 8) with still good quality can be achieved with a downscaling to 1280x720 and quantization with crf 26. Depending on the quantity of videos to be stored, it might be necessary to further reduce the data volume. A realistic scenario could be to use a hierarchical archiving strategy where the most recent and/or most important videos are first encoded in a high quality and after a specified period of time are transcoded with a lower bitrate and a just sufficient semantic quality for long-term archiving. Our study shows that this minimum required quality can be achieved with a quite low technical quality. Even the worst test condition in our assessment was far from being rated as insufficient although it only consumes 1.4 Mb/s (a reduction by a factor of 14).

V. Conclusions and Future Work

In this paper, we investigate the correlation of technical quality of laparoscopic videos and semantic quality subjectively judged by medical domain experts. We conducted a subjective quality assessment study with 37 participants to answer two research questions and conclude basic encoding recommendations for different scenarios. The main findings of our study are: (1) It is possible to compress laparoscopic videos to a much higher extent than in current practice without subjective quality degradation. (2) The semantic quality is still good enough with a low technical video quality with significantly reduced bitrate.

There is a large potential for future work in this field. As a next step, we plan to examine the correlation of objective quality metrics (e.g., PSNR, SSIM, VQM) with the MOS. Furthermore, similar studies should be conducted with low bitrate test conditions to investigate the actual boundary between just sufficient semantic quality and insufficient quality. In this context, it would also be interesting to examine the effects of the temporal resolution (frame rate) which is the third primary encoding parameter besides resolution and quantization. Recently, the new coding standard H.265/HEVC (high efficiency video coding) [8] has been introduced and presumably will establish as state-of-the-art in the next years. It is supposed to reduce the bitrate by a factor of 2 for equal subjective quality. It should be investigated if this is also the case for this domain. In consideration of the rapidly increasing volume of video data in the laparoscopy domain intelligent storage strategies based on our study should be developed and evaluated in practice. In this context it would also be important to initiate standardization processes, particularly with regard to future legal requirements.

REFERENCES