SUBJECTIVE MPEG2 COMPRESSED VIDEO QUALITY ASSESSMENT: APPLICATION TO
TELE-SURGERY

Nedia NOURI 1, Denis ABRAHAM 1, Jean-Marie MOUREAUX 2, Michel DUFUAT 1,
Jacques HUBERT 1, Manuela PEREZ 3

1 Centre de Recherche en Automatique de Nancy, Nancy-Université, CNRS
2, avenue de la forêt de Haye, 54516 Vandœuvre-lès-Nancy, France
2 Centre de Recherche en Automatique de Nancy, Nancy-Université, CNRS
Faculté des Sciences, BP 70239, 54506 Vandœuvre-lès-Nancy Cedex, France
3 INSERM, U947, CHU Nancy Brabois
Rue du Morvan, 54511 Vandœuvre-lès-Nancy, France

ABSTRACT
The digital revolution in medical environment speeds up
development of remote Robotic-Assisted Surgery and
consequently the transmission of medical numerical data
such as pictures or videos becomes possible. However,
video transmission requires significant bandwidth
and high compression ratios, only accessible with lossy
compression. Therefore research effort has been focussed on
video compression algorithms such as MPEG-2. In this
paper, we are interested in determining compression
thresholds and associated bit-rates which are acceptable with
respect to the quality required in the field of medical video
transmission. To evaluate MPEG-2 compressed medical
video quality, we performed a subjective assessment test
with a panel of human observers (experienced surgeons)
using a DSCQS (Double-Stimuli Continuous Quality Scale)
protocol derived from the International Telecommunication
Union recommendations (ITU-R BT-500-11). Promising
results estimate that 3 Mbits/s could be sufficient
(compression ratio around threshold compression level
around 90:1 compared to the original 270 Mbits/s) as far as
perceived quality is concerned.

Index Terms— video coding, Tele-surgery, biomedical
image processing, image quality assessment

1. INTRODUCTION
Progress of robotic-assisted surgical techniques allows today
mini-invasive surgery to be more accurate, providing
benefits to surgeons and patients. Furthermore, it opens the
field of remote surgery, as illustrated by the famous
experimentation “Operation Lindbergh” in 2001[1].
Currently, a surgeon operates while seated at a console
viewing the surgical field some meters away from his
patient (in the operating theatre) or at a distance (another
room), the data being transmitted from surgeon to robot via
a fully dedicated local network.

On the other hand, numerical data transmission across
long distances on a day-to-day basis remains a difficult
problem, as it requires significant additional bandwidth and
network resources. Hence, the transmission of medical
data such as video is feasible only if the amount of data is
significantly reduced, especially in the promising medical
class of Tele-surgery or Telementoring, where the volume
of data involved is huge. An appropriate solution to reduce
the amount of data could be found among the various
existing lossy compression schemes, and especially by using
the well-known MPEG-2 [2, 3] standard. As most video
coding techniques, MPEG-2 operates by removing existing
temporal, spatial, subjective as well as statistical
redundancies, by taking into account properties of the
Human Visual System.

But nowadays, compression methods used in medical
applications are most of the time lossless methods in order
to preserve the data integrity. Indeed, for example, by lossy
compressing medical video at high compression ratios,
compression artifacts can occur and disturb the progress of
surgical operation by decreasing significantly the quality
level of the video. However, it has been shown that a
balance between data compression and data fidelity could be
achieved in medical applications (see for example [4, 5]).

Thus, effective lossy compression represents an
important challenge for the future in the medical field, as, on
the one hand, only lossy techniques yield a significant
reduction of the amount of information, and on the other
hand, the involved distortion must not affect the goal of the
application.

To assess the effect of compression on perceived image
quality, objective measures like Peak Signal to Noise Ratio
(PSNR) are generally used in the literature, for sake of
simplicity. However, it is known that the PSNR is a global criterion which is not sufficient to measure the quality of processed images or videos, especially in medical applications where the evaluation should be done by experienced doctors over a large data set. Unfortunately, such studies are very expensive in terms of time and human means and thus are rare in the literature.

Here we propose to carry out such a study in order to determine the optimum range of bit-rates to be used in the Tele-surgery context, by using MPEG2 compressed videos. Our main contributions are threefold. First, as said above, we explore the field of lossy compression for Tele-surgery. Second, our study is based on the subjective evaluation of MPEG2 compression impact achieved by experienced surgeons rather than on PSNR evaluation. Finally, we propose to adapt the Double-Stimuli Continuous Quality Scale (DSCQS) from the ITU-R BT-500-11 recommendations [6, 7] to the context of Tele-surgery, which has not yet been used at our knowledge in this domain. We show that lossy compression can be achieved in the framework of our application until 90:1 compression ratio without significant decrease of the video quality, according to experienced surgeons.

This paper is organized as follows. In section 2 we introduce the subjective quality assessment methodology and its adaptation to Tele-surgery context. Section 3 shows the experimental results, whereas conclusions and perspectives are given in section 4.

2. SUBJECTIVE QUALITY ASSESSMENT

The general test viewing conditions for subjective assessment of the quality of television pictures are given in ITU-R BT.500-11 recommendations. To conduct appropriate subjective assessments in the context of Tele-surgery, it is first necessary to select from different available options, those that best suit the objectives and circumstances of the assessment problem. Some of test methods are double stimuli where viewers rate the quality or change in quality between two video streams (reference and impaired). DSCQS method is claimed to be less sensitive to context: the subjective ratings are less influenced by the severity and the order of impairments within the test session. Thus, we choose this method for our quality assessment trial.

In next sections, we describe the general viewing conditions and the DSCQS method that was used for our subjective quality assessment of MPEG2 compressed video sequences trial.

2.1. General viewing and scoring conditions

Several test procedures for subjective quality evaluation are defined in ITU-R BT.500-11. Among them, a commonly used procedure is the DSCQS method in which an assessor is presented with a pair of video sequences A and B, one after the other (see FIG. 1), and asked to give A and B a ‘quality score’ by marking on a continuous line. This line is divided into five equal lengths intervals which correspond to the ITU-R five-point quality scale (‘Excellent’ to ‘Bad’), as illustrated in FIG. 2. In a typical test session, the assessor is shown a series of pairs of sequences and is asked to grade each pair. Within each pair of sequences, one is an unimpaired ‘reference’ i.e. an original sequence and the other is the same sequence, modified by a process under test, here lossy compression. The selection of which sequence is ‘A’ and which is ‘B’ is randomized during the test session.

<table>
<thead>
<tr>
<th>Seq.</th>
<th>Bit-rate (Mbits/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.02 4.02 5.04 6 8.04</td>
</tr>
<tr>
<td>2</td>
<td>1.5  2.56 4.5 5.52 7.5</td>
</tr>
<tr>
<td>3</td>
<td>1.74 2.76 3.78 4.74 5.76 6.78 7.74</td>
</tr>
<tr>
<td>4</td>
<td>1.2  2.04 2.28 3.24 4.25 5.28 6.24 7.2</td>
</tr>
</tbody>
</table>

The scores obtained by each sequence on the continuous quality scale during assessment at several bit-rates (from 1.02 Mbits/s to 7.2 Mbits/s – see TABLE I), were processed as the difference in judgment between reference and impaired sequences. (See FIG. 2).
2.2. Analysis and interpretation of the results

A test consists of \( L \) presentations (i.e. \( L \) pairs), each presentation belonging to a set of \( J \) test conditions (i.e. bit-rates). Furthermore, some of the test conditions are repeated \( K \) times, in order to check the reliability level of each assessor. Note that \( K < L \) presentations are really effective. In this study, \( L = 38 \), \( J = 25 \), \( K = 28 \) and \( R = 3 \).

The first step of the analysis of the results is the calculation of the mean score \( \bar{u}_{jk} \) for each of the presentations.

\[
\bar{u}_{jk} = \frac{1}{N_{\text{obs}}} \sum_{i=1}^{N_{\text{obs}}} u_{ijk}
\]

with \( N_{\text{obs}} \), the number of assessors. Mean scores are represented in figures 3, 4, 5 as a function of the bit-rate.

All mean scores have an associated 95\% confidence interval which is derived from the standard deviation \( \sigma_{jk} \) and size of each sample.

\[
\left[ \bar{u}_{jk} - 1.96 \left( \frac{\sigma_{jk}}{\sqrt{N_{\text{obs}}} \cdot I} \right), \bar{u}_{jk} + 1.96 \left( \frac{\sigma_{jk}}{\sqrt{N_{\text{obs}}} \cdot I} \right) \right]
\]

The standard deviation \( \sigma_{jk} \) for each presentation is given by:

\[
\sigma_{jk} = \sqrt{\frac{N_{\text{obs}}}{\sum_{i=1}^{N_{\text{obs}}} \left( \bar{u}_{ijk} - \bar{u}_{jk} \right)^2}}
\]

3. EXPERIMENTAL RESULTS

In this study, seven assessors (expert surgeons) were used for the subjective test. Usually, recommendations invite to perform the subjective test by getting involved either more than fifteen non expert viewers or more than four expert viewers. Considering that, a representative panel of surgeons supposes a set of viewers that are trained to work on videos; it is then natural to consider all of them as experts. The choice of using seven observers is then considered as relevant. Furthermore, four video sequences, representative of four typical situations faced by robotic-assisted surgeons in standard conditions, were chosen.

Besides, ITU-R recommendations have been written in order to reproduce real life conditions for the subjective tests. Unfortunately, ITU-R has not standardized any specific protocol for robotic-assisted surgery yet. So, to follow the ITU recommendations for keeping test conditions as close as possible to surgery real environment, we decided that our test protocol should be performed in a real operating theatre. Any aspect of the test has been thought in that direction. Thus, video sequences were displayed to observers thanks to the viewing facilities of the robotic-assisted console. Moreover, voting tools were delivered in compliance with the operating theatre typical conditions.

The original sequences have been provided by a digital component CCD camera (endoscope) in PAL format digitized in 4:2:2 format. The sequences were processed by a MPEG-2 hardware encoder at variable bit-rates (see Table 1) with a GOP size fixed at 12, then were transmitted over operating theatre local network and were displayed on the robot viewing console.

Individual observers’ scores were stored in a test database. Furthermore, ITU-R recommendations describe the statistical procedure to process the database aiming at deriving interesting results. The difference between reference and impaired sequence (on the continuous quality scale) is considered when processing the test database. The lower the value, the less the impairment is and the less critical the MPEG-2 sequence is.

When exploiting the scores provided during the test, one observer’s behavior appeared to be atypical as if he was following his own score scale. To keep coherent the subjective test database, this observer was automatically rejected from the observer list.

Some noticeably results are following. When exploiting the test database, some sequences may appear to be not sensitive to the variation of the technical degradation introduced (under test). Such cases exist: the concerned sequences are then considered as non critical in the sense that these do not reveal sufficient loss of quality linked to the introduced degradation, whatever the level of degradation is. Sequences 2 and 3 appeared to be non critical as far a bit-rate variation is concerned, up to the lowest value of bit-rate: 1.5 Mbits/s.

Hopefully, perceived quality sequences 1 and 4 vary significantly with variation of bit-rates. The great impact on perceived quality due to bitrates reduction is really not comparable to a linear relationship. In fact, two levels of quality are revealed by bit-rate reduction, as if the decrease of bit-rates was not perceivable up to a certain limit and as
below this limit, the quality becomes very poor. In such promising situation, there are good reasons to identify a model as an approximation of the cloud of points (bit-rate vs. Mean score for perceived quality). A regression curve is identified, thanks to least square approach. In addition, coefficient of determination corresponding to the dispersion of points compared to the curve provides very promising results: \( R^2=0.6245 \) for sequence 1 and \( R^2=0.9117 \) for sequence 4. The highest the determination coefficient in the range of 0 to 1 is (considering human subjectivity, 1 is never seen), the more efficient the model is as well as quality of results and derived conclusions. In this situation, we are in position to identify, with a good precision, the threshold, above which no loss of quality is perceived and below it only a poor quality can be expected.

Conclusions of this study are then very strong: above 3.2 Mbits/s, no surgeon perceived any loss of quality due to bit-rate reduction. Considering that the same video requested initially 270 Mbits/s to be transmitted, this means that a compression ratio of around 90:1 can be adopted in the field of robotic-assisted surgery, without being exposed to any medical risk due compression algorithm side impact. In the field of medical applications, in general and more precisely in a surgery application, this very promising conclusion is of high importance, unlocks one of the technical locks to transmission of medical video and opens doors to distant-teaching, distant-mentoring and soon distant-surgery.

One of the main results of this study is that, not only using lossy compression for medical video transmission seems possible, but also one can determine a threshold above which no surgeon perceives any loss of quality. In the framework of our application, this threshold is around 3 Mbits/s. In fact, it is relevant to add a safety margin of 20% in order to anticipate critical or complex sequences. This result could be considered of high importance since it was established thanks to a protocol and procedure of test derived from the famous ITU-R BT.500-11, internationally recognized and especially adapted here for robotic-assisted surgery. The subjective assessment test was performed in real surgery conditions: in a surgery theater, with a panel composed of surgeons.

Considering the so promising conclusions, future work, will go on with the same application and will investigate the use of other encoding techniques such as H.264.