Subjective Quality Assessment of the Impact of Buffer Size in Fine-Grain Parallel Video Encoding

Pablo MONTERO[1], *Ladislav POLÁK*[2], *Javier TAIBO*[1], *Tomáš KRATOCHVÍL*[2]

[1]MADS Group, University of A Coruña, CITIC-research building, Campus de Elviña s/n, 15071 A Coruña, Spain
[2]Dept. of Radio Electronics, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic

pmontm@gmail.com, xpolak18@stud.feec.vutbr.cz, jtaibo@udc.es

Abstract. *Fine-Grain parallelism is essential for real-time video encoding performance. This usually implies setting a fixed buffer size for each encoded block. The choice of this parameter is critical for both performance and hardware cost. In this paper we analyze the impact of buffer size on image subjective quality, and its relation with other encoding parameters. We explore the consequences on visual quality, when minimizing buffer size to the point of causing the discard of quantized coefficients for highest frequencies. Finally, we propose some guidelines for the choice of buffer size, that has proven to be heavily dependent, in addition to other parameters, on the type of sequence being encoded. These guidelines are useful for the design of efficient real-time encoders, both hardware and software.*

Keywords

Parallel MPEG-2, ITU-R BT.500, subjective test, video quality, ACR, MOS, PSNR.

1. Introduction

Video encoding has been an active research field for decades, with important applications in the industry. The most widespread encoding schemes, such as MPEG-2 [1] or H.264 [2], can be divided in separate tasks, such as DCT (Discrete Cosine Transform), quantization, motion estimation, entropy encoding, rate control and others. The pixels in each frame are grouped in blocks. Thanks to this, parallel processing of these tasks is possible. This property has been studied in [4] among other works. Parallelism is especially important in real-time video encoding, where time is critical.

Entropy encoding is one of the most difficult tasks to parallelize, as it has a serial nature due to dependencies among processed data. However, there are some proposals for its parallelization [5], [6]. The result of this entropy encoding is a variable length bit stream. Without prior knowledge about the size of the bit stream for each block and processing the blocks in parallel, fixed-size buffers have to be allocated for the resulting bit streams.

The size of these buffers is an important decision, that has not received too much attention in previous works. This decision is sometimes addressed by overestimating to a buffer size (large enough) that will not produce overflows; in [7] a 2048 bits buffer per macroblock (16×16 pixel) is used, considered enough for 99.99% of the possible scenarios. In [3] a maximum of 416 bits is reserved per 4×4 block.

An excessive buffer size may have the consequence of increasing the encoding time as it increases the transfer times [6]. This encoding time is critical in some scenarios, such as the real-time massive encoding of video streams. Moreover, in hardware implementations [8] memory constraints are more severe and the decision about buffer size is absolutely critical to minimize hardware cost.

On the other hand, a low buffer size increases the risks of buffer overflow, which, when not controlled, produces an invalid bitstream and when controlled introduces additional noise, sometimes causing noticeable artifacts in the image. A simple procedure in the presence of buffer overflow consists in discarding the quantized coefficients, starting from the highest frequencies until there is no overflow. This minimizes the artifacts, since the human visual system is less susceptible to high frequency details [9]-[11]. This decreased visual impact allows for some reasonable "trade-off" or "risk taking", when selecting buffer size, where otherwise the artifact would be unacceptable. This step can be done in most standards from MPEG-2 to H.264, since all are based on DCT and zigzag scan. It is important to note that the artifacts caused by this coefficient discard have a different nature than the ones produced because of higher quantization steps: the former simply removes some coefficients without changing the rest and the latter uniformly reduces all coefficients until some disappear or take less space when entropy encoded.

The quantization parameter, responsible for the lossy part of video encoding, affects the bitstream size for each block and its visual quality. Quantization and buffer overflow affect image quality in ways different to each other. Increasing the quantizer will reduce image quality, but also bitstream size. Therefore, a smaller buffer is needed and/or buffer overflow danger is reduced. If the bitstream size reduction due to quantization avoids a buffer overflow, the decrease in quality by a stronger quantizer can actually be

a quality improvement, when compared with a smaller quantizer that would produce a buffer overflow. This effect is show in Fig. 2.

If a buffer overflow is detected, re-encoding with a stronger quantizer is a better solution in terms of visual quality. However, for the real-time encoding scenarios this is not a good option, since it would seriously harm performance.

The target at this point is to answer the following questions: What is the optimal buffer size? How should the buffer size be chosen? To answer these questions, we should analyze the combined impact of both buffer size and quantizer over image quality. This impact may be measured with objective methods, such as PSNR. On the other hand, our previous experience [6], [14] showed that for a similar PSNR, the same sequence in the presence of buffer overflow produced more noticeable artifacts and therefore worse overall quality, compared with higher quantizer without overflow. Reiter and Korhonen [15] showed how PSNR is less reliable when comparing distortions of different types.

This fact motivated the subjective quality assessment experiment that is presented in this work. We analyzed the subjective quality for several video sequences, encoded with different parameters. We studied the impact of buffer size in relation with the quantizer value. The experiments were based on the recommendation ITU-R BT.500 [19]. We centered our study on real-time generated synthetic video (i.e. computer generated images), because we are interested in real-time encoding for interactive applications. However, the results can be extrapolated to natural video based on the high percentage of this kind of content in some sequences.

The paper is organized as follows. Section 2 briefly presents a short overview of actual related studies in the area of subjective tests. Section 3 describes the experimental design. Furthermore, the video material used for the testing, together with the definition of main features and parameters, is presented and described. Section 4 describes how the design was implemented. Section 5 provides an analysis of the obtained results. Finally, the paper concludes in Section 6.

2. Background of the Subjective Tests

In the recent years, the video processing community has been largely interested in the quality aspects of the video content, delivered to the user. In the effort to provide more and more different TV program material with high quality to a higher number of consumers, there is an increased amount of information being transferred over the same limited bandwidth [16]. This is also the reason why it is important to provide and study the quality of video.

Video quality can be determined using objective measurements [17], [18] and by subjective assessments [19], [20]. The most common methods, which are used to measure subjective quality of videos, are also recommended by the ITU [19]. Nowadays, several types of video quality assessment methodologies exist. They are defined in the ITU-R BT.500 [19].

Several studies on subjective video quality assessment have been carried out. These studies are mainly focused on exploring the different methodologies for the subjective assessment of the quality of television pictures and multimedia applications (e.g. [21], [22]), on the analysis of temporal effects in quality assessment of HD (High Definition) video, when different video codec were used (e.g. [23]-[26]) and on the quality assessment of very perspective 3D video quality [27]. Inspired by the mentioned papers, we decided to analyze the image quality of video stream encoded as described above. The details will be outlined below.

3. Experimental Design

We used the GPU parallel MPEG-2 implementation described in [14] for our experiment. Only intra-frames (GOP size 1) were used, to prevent the propagation of errors and negate the influence of motion estimation algorithms.

The purpose of the experiment was to provide a better understanding on how buffer size affects quality and how it relates to other parameters. With this in mind, we identified the most important parameters that influence video quality, and among them, selected different test values which are shown below. We created a test clip for each selected combination of parameters, and each test subject watched all these clips. The parameters considered for experimentation were the following:

- **Buffer size (σ):** The buffer size in bits is for a 8×8 block. Values are 40, 104, 328. These specific sizes were chosen based on their impact on performance and visual quality. Sizes below 40 produce very small benefits. Sizes over 328 almost never produce buffer overflow. While more levels of this parameter would help modeling the behavior, it is out of the scope of this work.
- **Quantizer (q):** Values are 2, 5 and 9. Higher quantization levels have a quickly decreasing probability of producing buffer overflow at the minimum chosen buffer size, 40.
- **Sequence:** Different sequences were chosen out of the hypothesis that both σ and q would affect the quality in different ways depending on the content of the video. Then, a worst case or best case could be identified, avoiding the risk of using a most favorable sequence for the experiment. We chose five video sequences with different characteristics (see Fig. 1):
  - **SciFi:** Slow fly through a starship corridor. Combined panning and zoom in/out. Complex spatial features.
  - **Magazine:** Sliding and zooming magazine pages containing text and still images. Text is known

to produce highly visible artifacts when processed by non specific lossy image encoding algorithms, such as MPEG-2 or H.264.

- **Dragon:** Character animation test over an empty background. A little baby dragon runs round and round. Fast motion, no complex spatial features, round figures, lighting, textures.
- **Megazapper:** Interactive TV channel browser. Live TV streams are arranged in an animated 3D layout. High details, complex spatial features, natural video.
- **Elephant:** A clip taken from Elephants Dream animated short movie. State of the art 3D animation. Standard sequence that can be used for comparison.

- **Test subject:** The behavior of each subject was considered to detect differences in how each person reacts to the different parameters of the sequence.
- **Video resolution:** Considered resolutions were PAL (720×576), 720p and 1080p. As the σ parameter defines a per-block buffer size, the resolution establishes the space screen of each block. For the sake of limiting the amount of test clips that each subject must watch, we finally performed the tests with 1920x1080 (1080p) resolution.

In order to further decrease the amount of sequences, some combinations of σ and q were discarded, as we already knew that they do not produce overflow or produce an excessive amount of it. The selected combinations of (q,σ) were: $(2,104)$, $(2,328)$, $(5,40)$, $(5,104)$, $(9,40)$ and $(9,104)$.

One test clip was created for each video sequence, σ, and q combination. The order of presentation of these test clips was grouped by sequences and random for the other two parameters. Once generated this random list, all subjects were presented the same succession of test clips. We grouped the test clips for each sequence to ease the evaluation process, as changing sequences would difficult the appreciation of more subtle changes of q and σ among the same sequence. The final amount of test clips was 30: 6 combinations (q,σ) for 5 sequences. The length of the experiment for each test subject was 10 minutes.

(a) SciFi **(b)** Magazine **(c)** Dragon **(d)** Megazapper **(e)** Elephant

Fig. 1. Screenshots of the sequences used in the experiment.

(a) quantizer = 5, $\sigma = 104$ **(b)** quantizer = 5, $\sigma = 40$ **(c)** quantizer = 9, $\sigma = 40$

Fig. 2. σ = buffer size in bits. It is shown how increasing the quantization step size can produce visual improvements in case of buffer overflow (Fig. 2b - Fig. 2c).

4. Test Session

This section briefly describes the implementation of the designed experiment. The hardware and software solutions, used for presenting the test clips and method for ratings of the quality of video sequences will be outlined.

4.1 Participants

In our experiment, overall thirty people (17 males and 13 females) have participated, recruited from university workers. The average age of participants is equal to 33. Only five of them had expertise in video processing and quality assessment. None of them had previously participated in similar tests. More than 50 percent of the viewers (concretely 18) had glasses, but only for reading. On the other hand, this, so called vision defect, had no significant effect on the results, because all observers were tested for color blindness (see 4.2). The whole experiment was realized in the usability laboratory, Living Lab Galicia [28], at the CITIC research center, based in A Coruña (Spain).

4.2 Color Vision Test

Firstly, all participants were tested for visual acuity and color blindness prior to the test using the Ishihara chart. For this purpose an appropriate application was created in MATLAB with GUI (Graphical User Interface). This application allows for testing the color blindness of the user. The user,

after starting this application, will see a sequence of ten different types of Ishihara chart. For each one, he or she must type the number, which is shown in the picture. At the end, from the obtained results, the average rate is calculated and the results are saved into a text file. In case of deep interest on the full version of this application, the code is available on request.

In this test everybody achieved a high success (the average rate of success was 91 %). Therefore, the scores given by thirty users were analyzed.

4.3 Test Setup

The whole subjective video quality assessment experiment was conducted in a special test room, with simulated home conditions. More precisely, the room had controlled lighting and for the observers a comfortable couch for the sitting was prepared. The ambient light in the room was 300 lux. For the ambient light, fluorescent lamps were used.

The presentation of short video sequences was done on a personal computer. The output of this PC was brought via digital interface to a calibrated 46" Samsung SmartTV. The luminance of the LED display was adjusted to 200 cd/m^2. Following the ITU-R BT.500 recommendation [19], the viewing distance for all observers was approximately between 3.5 H and 4.5 H (depending on viewer height and pose), H being the physical height of the picture. The viewing conditions for the observers were ensured, as they are recommended in [19], [20]. The tests were performed in a pipeline. More precisely, when a person completed the color vision test, he or she came and did the assessment of the quality of video sequences while the next person started the color vision test.

As the test setup requires collecting the users' ratings over time, specific user interface (hardware and software) or a special questionnaire needs to be used. Nowadays, both of them are frequently used in the field in quality assessment of video. For example, in [16], [23] and [24], for the observers, a slider interface (hardware and software) was developed, based on the guidelines, presented in [19]. The interface uses a continuous quality scale with numeric values, e.g. reaching from 0 (worst quality) to 100 (best quality). In [27], a simple questionnaire is used for the recording of the score from observers.

In our experiment, we decided to combine both methods. As mentioned above, overall thirty people participated in the experiment. Therefore, we divided the people into two groups. Both groups had fifteen people (as recommended in [20]). First group provided their quality ratings electronically, using a computer mouse wheel. For this purpose, an application was developed, which can respond to the decision of the observers. The "PC" slider interface was also visible at the bottom of the TV screen and of course, it uses a continuous quality scale, as mentioned above. The second group used a prepared questionnaire into which their quality ratings were manually recorded. The details of assessment methodology and rating scales that were used in our experiment will be outlined in the next subsection.

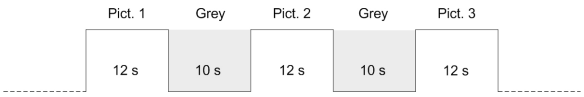

Fig. 3. Stimulus presentation in the ACR method.

4.4 Assessment Methodology and Rating Scales

All experiments were conducted using a well known SS (Single-Stimulus) presentation, as used in ACR (Absolute Category Rating) method [20]. This method is generally used [16], [24], because it is the simplest and the fastest method. In our tests we used the SS method, but the presentation procedure was the same as it is in the ACR method. The ACR method is a category judgment, where test sequences are presented one at a time and are rated independently on a category scale. The time pattern for the stimulus presentation, which was used in our experiment, is shown in Fig. 3.

As shown in Fig. 3, the video sequences were presented one after another (Pict.1, Pict.2), with a 10 second gray screen image between them. The duration of each video sequence is equal to 12 seconds. Of course, each video sequence is presented one at a time and rated individually, according to the time pattern.

The investigation of the best rating scale is one of the most important elements in the field of subjective video quality assessment. Quan Huynh-Thu et al. [16] explored the effect of rating scales on the subjective scores, collected using a given stimulus pattern presentation. They made a direct comparison between four different scales, which are included in existing international standards. Surprisingly, the data obtained show no overall statistical differences between the different scales.

For our experiment, presented in this paper, we decided to use the 11-point continuous scale (in both groups). As briefly mentioned in [16], this scale is mainly used in the field of audio quality. On the other hand, it has been also widely documented and applied in the field of video and picture quality assessment. However, this type of scale is used for comparison to an original signal, indicating fact that this scale was originally used in the DS (Double-Stimulus) approach [19]. However, in this paper we used the SS method, where comparison of the video sequence to original is not possible. Therefore, the definition of the two extreme points was slightly modified: number 0 represented the worst possible quality and number 10 represented the best possible quality, as it was solved and applied in [16]. These facts for the participants were also clearly explained.The scales, used throughout our experiment, are shown in Fig. 4 and Fig. 5.

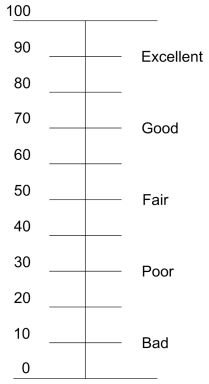

Fig. 4. Rating scale, used in the first experiment: 11-point continuous scale.

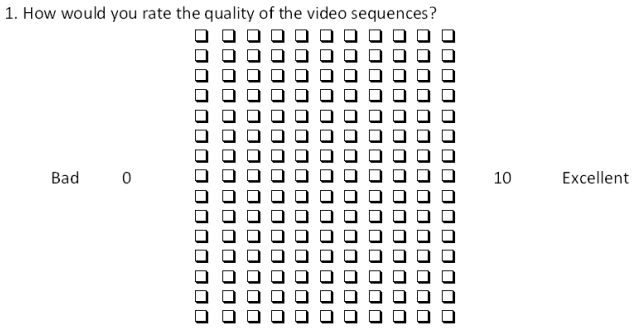

Fig. 5. Rating scale, used in the second experiments: Questionnaire.

5. Data Analysis

We studied the data from two points of view: the subjective quality evaluation and their temporal aspects. For the subjective quality evaluation, the input data is formed by the slider ratings and the ratings in the written form. The slider and form data have different nature and can not be subject to the same methods of analysis. They were studied independently and compared at the end.

5.1 Analysis of the Slider Data

For the subjective quality analysis of the slider data, only the last input of each user on each test clip was used, as it would represent their final decision about the quality of the clip. On all statistical tests performed, the used confidence level was 0.95.

5.1.1 Global Data Analysis

The first step was checking the validity of the data for the application of statistical inference.

- The data gathered with the slider device is considered of normal distribution on a Shapiro-Wilks test of normality [32].

- Homoscedasticity test between sequences is also positive in a Levene's test [33].

- Randomness and independence can be assumed based on the random order of presentation of test clips. The influence of the weariness of the observer is minimized by having a overall short test.

After these verifications, we proceeded to the in-depth analysis of the data.

5.1.2 Analysis of Variance

This analysis can reveal whether there exist real differences in subjective quality between test subjects, sequences, quantization levels and so on, when taking into account the inherent randomness that emerges in every experiment. It is perhaps more important, since it is more difficult to observe without an experiment of this kind, the underlying relationship between pairs of parameters (interactions).

- **Sequence:** There is a statistically significant difference on the evaluations between different sequences. This is not entirely unexpected, as the complexity of the sequence is determinant in the influence of q and σ parameters. It is also very difficult for the subject to, as instructed, abstract itself from the content of the sequences and evaluate only the visual quality of each sequence. It is usual that more aesthetically pleasing scenes are rated higher. The content can also influence the perception in other ways, like sharp edges, textures and aliasing.

- **Quantization:** The influence of the quantization parameter is statistically significant. The test subjects performed good in assessing the visual differences introduced by the quantization level, besides being relatively small (2,5,9).

- **Buffer size (σ):** The influence of the σ parameter is statistically significant. This experiment proved that the buffer sizes studied produce a difference in the visual quality of the encoded video. This is one of the main focuses of the experiment and it will be more closely studied in Section 5.1.3.

- **Test subject:** Each test subject valued differently, what can be attributed to personal preferences. This fact was taken into account and can have some implications, especially in the interactions of other factor with the test subject.

The interaction between factors were also considered of great interest:

- **Sequence - quantization:** Interaction between the sequence type and the quantization proves that the quantization degrades each scene in a different way. This interaction is widely known, caused by how some spatial and temporal features are degraded when quantizing. It further validates the experiment data, as the test subjects were able to evaluate the quality accurately enough to reproduce this interaction.

- **Sequence - σ:** The interaction between sequence type and σ level is significant, explained by the same reasons as the interaction between sequence and quantization.

- **Quantization - σ:** This interaction was proven significant in our experiment. The existence of an interaction between the quantization and σ parameters is of special importance for us, as it signals athe behavior: The influence of each parameter in the perceived quality changes with the level of the other. A closer study is given in Section 5.1.3.

- **Test subject - quantization:** There is no interaction between these factors. Knowing that each test subject responds to each level of quantization in the same way contributes to the overall validity of the experiment, as it is an expected behavior.

- **Test subject - σ:** There is no interaction, explained by the same reasons as the previous interaction.

- **Sequence - test subject:** There is interaction between these two factors. This means that each subject evaluates each sequence in a different way. We attribute this to personal preferences in the sequence content, that the test subjects can not isolate from, when evaluating visual quality.

- **Sequence - quantization- σ:** This proved that the relationship between quantization and σ is of the most complex nature: it changes with the content.

The rest of the combinations were not significant. They are interactions, where the test subject factor is involved, and the interpretation of the lack of significance is similar to the other not significant interactions with test subject. A summary of the results is given in Tab. 1.

5.1.3 Interaction q - σ

The σ parameter dictates the buffer size and the quantization level q influences how much buffer space is used. When the quantization level is low, the bitstreams of the blocks are bigger, so the importance of the σ parameter is increased. On the other hand, as the quantization level increases and the block bitstreams become smaller, the σ parameter must be extremely low to produce visible artifacts.

This is proven in the experimental results, but also the quantitative relationship is much importance: "How much σ is too much?" For any given quantization level there is a point from where increasing σ ceases to produce benefits in the perceived quality. The experimental results told us that this point changes with the content of the video, but is the same for all participants. For the studied combinations of quantization and σ parameters, we can compute whether there is a statistically significant difference between them using a Tukey's range test [31]. We now proceed to explain the most relevant results:

	p-value
sigma	0.0000
quant	0.0003
sequence	0.0000
test_subject	0.0000
sigma:quant	0.0002
sigma:sequence	0.0009
quant:sequence	0.0000
sigma:test_subject	0.0678
quant:test_subject	0.1806
sequence:test_subject	0.0005
sigma:quant:sequence	0.0005
sigma:quant:test_subject	0.3439
sigma:sequence:test_subject	0.6204
quant:sequence:test_subject	0.6001
sigma:quant:sequence:test_subject	0.7278

Tab. 1. ANOVA p-values. Shaded in gray are the non significative interactions.

For $q = 9$ there is no difference between the studied σ levels: $\sigma = 40$ and $\sigma = 104$. For the rest other studied quantization levels: $q = 2$ and $q = 5$ there is a difference in the perceived quality for the different levels of σ. This tells us that the σ used should be higher than 40 or 104, depending on the quantization level.

The fact that the influence of test subjects in the relationship between σ and q is proven nonexistent opens the possibility of a better experimental approach could be used, where each test subject would not have to watch and rate all the different sequences but a specific subset of them. Therefore, each sequence would receive the same amount of evaluations but the burden of the test subjects would be lessened. Also more combinations of parameters could be measured.

5.1.4 Effects of σ on Bitrate

The σ parameter also controls the size of the block, effectively influencing the bitrate of the sequence. Can σ produce benefits in bitrate while not deteriorating visual quality? To provide an answer for this question, we can use the already gathered subjective evaluation data, and analyze it from the bitrate point of view. Previous analysis showed that there are combinations for (q,σ) that produce the same subjective quality than others. Examples of these combinations are $(2,104) \approx (5,104)$ and $(5,40) \approx (9,104)$. Given these combinations, we proceeded to check the bitrates of these sequences. At the same quality, the bitrate goes from 13% to 23% higher in the cases that the σ is underestimated, i.e the bitrate of sequence with $(q,104)$ is $13-23\%$ higher than the same sequence with $(5,104)$ and produces that same subjective quality. This result discards further study in the selection of σ for bitrate purposes.

5.2 Influence of Sequence Type in q, σ and q-σ Interaction

As a result of the statistical significance of the interactions sequence-q, sequence-σ, and sequence-q-σ, a closer look in this relationship is deemed of interest. For each se-

quence type, the differences produced in the responses were measured. For the studied quantization levels, we compare the mean scores for the σ parameters in Fig. 6.

The sequence that consistently maximizes the difference in subjective quality between σ levels is Magazine, a scene consisting mostly of text, that produces edges that make the difference in σ highly noticeable. On the more "natural-like" scenes (Megazapper and Elephant), the difference fades quickly on quantization level 9, while on the rest of the sequences the difference stays stable. As the influence of the sequence content is proven relevant, and modeling the scene content could be very complex or even impracticable for real-time purposes, a scene containing the worst case scenario, the scene that maximizes the influence of σ, should be used for that experiment. Based on these results, the scene suggested for further study on the influence of σ is Magazine.

Fig. 7 shows how quality is related to q and σ. It can be seen that the sequence type is not only an additive factor, it actually changes the shape of these two relationships, though the effect is stronger on q than on σ.

Despite all sequences are synthetically generated, some of them include natural (i.e. acquired with a camera, as opposed to computer generated) video or pictures as textures. It is interesting to study the behavior of the different sequences as a function of "how synthetic they are". To sort these sequences, we considered features like the use of textures, texture contents, lighting, geometry complexity, sharp edges, or aliasing. These features exposed significant differences in the perceived quality for the different q and σ parameters. Our results showed that natural contents are more forgiving than synthetic contents in terms of subjective quality.

A good example of this are sharp edges, produced by synthetic geometry or rendered text. They produced highly noticeable artifacts even at relatively low quantization levels. This may be the reason that Magazine sequence has got the worst results, as the majority of the frame is composed of text. The best scored sequences were Megazapper and Elephants Dream. Megazapper is a 3D mosaic of natural videos. Elephants Dream is a synthetic scene with a high quality render that approaches it to the aspect of a natural scene, in comparison with the other sequences.

5.3 Analysis of Temporal Data

The temporal data were gathered to look for patterns in test subject scoring, such as points in the sequence that produced quick changes. Conclusions of this analysis, such as increasing test clip length, could be used to improve further research. A total of 7636 inputs were captured on the slider, with an average of 510 per test subject and 17 per sequence. The mean stabilization time, when the test subjects selects the final score on the slider, is 11.3 seconds out of the 12 that the sequence takes and 10 of the gray separator between sequences.

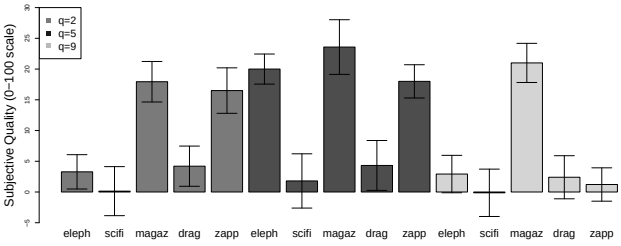

Fig. 6. Differences in sequence evaluations between the two σ values for each quantizer with 95 % confidence intervals. From left to right, $q = 2$, $q = 5$ and $q = 9$.

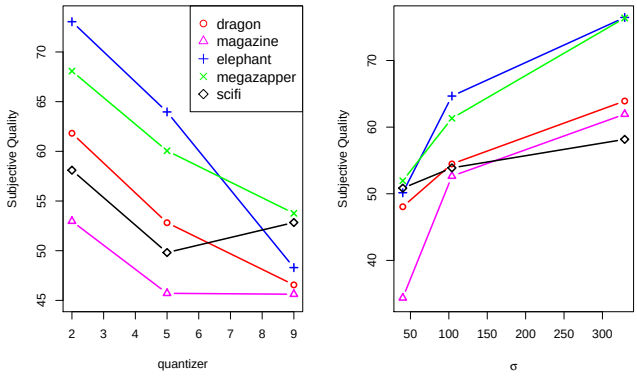

Fig. 7. Subjective quality for quantizer and σ.

An interesting question raised is whether the different parameters affect the stabilization time. Using a Kruskal-Wallis test [34] we can establish:

- There are differences in stabilization times between different sequence types. Some scenes required more time than others
- There are no differences in stabilization times between different σ values.
- There are no differences in stabilization times between different q values.

With these results we can interpret that the unique factor that defines the rating times is the sequence content and not the image quality, for the values parameters studied.

Only 1.1 % of the evaluations ran into the last second of the time provided, suggesting that the time for evaluation was enough.

When we studied the patterns, we observed that most were monotonically increasing or decreasing, with a very high variability between test subjects. This is also confirmed by the proven difference in the evaluation scores between test subjects. It is interesting that besides this wide diversity, there is no difference in how different sequences, quantization levels and σ values affect each test subject individually. They are all affected in the same way.

Fig. 8 shows some patterns for the same test clip by different test subjects.

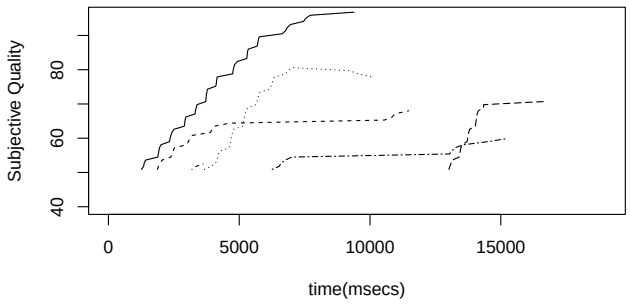

Fig. 8. Time pattern for 6 test subjects on the same test clip.

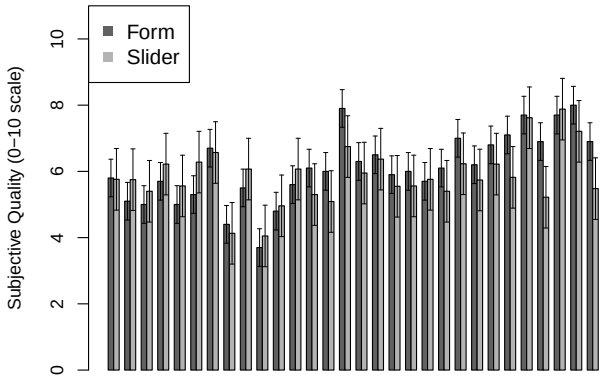

Fig. 9. Comparison of average evaluations between slider and form data with 95 %. confidence intervals.

5.4 Comparing Slider with Form Data

Both methods gave very similar results, in terms of means and difference between sequences, quantization levels and σ values, with a total Pearson correlation between means for each test clip of 0.76.

The mean score was 6.2 for the form data and 5.9 for the slider data. The main difference is between the standard deviations. It was 1.84 for the form data, while the deviation of slider data was 1.0, after scaling to the 1-10 scale. We believe that this is a natural consequence of limiting the selection of values for scoring to 10 discrete numbers.

Another interesting comparison is the number of mistakes, where a mistake is defined as: "for each sequence type, test clip A with the same quantizer as test clip B but greater σ scoring lower average". Only two such mistakes were made on each of the methods, and were located in the first sequence shown, this could be attributed to the "learning period" of the experiment.

Finally, Fig. 9 shows the comparison between form and slider average evaluation of the 30 test clips, unified to the 1-10 scale, for a quantitative illustration of the differences.

6. Conclusions

The buffer size used in entropy encoding is critical for real-time encoding, because of both hardware cost and encoding performance. Insufficient buffer size forces coefficient discard that causes visual artifacts. We analyzed the impact on visual quality of this buffer size (σ), reserved for each block of compressed video. For research and non-commercial use, the used video sequences can be retrieved from [30].

Following the findings of previous work, we performed a subjective quality assessment, as objective methods are not adequate in this case. From the experiment results, we can conclude that the studied buffer sizes affect perceived quality. We further analyzed the relation of the buffer size with other parameters that affect quality.

There were no significant differences among subjects in their response to different σ values. This means that all subjects change their scoring the same way in relation to σ value. This is an important conclusion, as subjects can be considered equivalent for the measurement of their response to changes in σ. This can drastically reduce the complexity of future experiments.

The experiment showed that the influence of σ on the perceived quality is reduced when the quantizer value increases. However, some sequences do not show this effect as clearly as others. This makes it difficult to predict the σ value as a function of the quantizer without further knowledge about the sequence to encode.

The use of σ to improve the encoding efficiency was discarded, as it gives worse subjective quality than adjusting the quantizer to achieve the same bitrate.

There was a good correlation between the results of the slider test and the written form, a fact that confirms the results of [16].

We found that only sequence affects the time that users need to score the test clip. No other of the studied parameters were found significant.

The data gathered in our experiment were insufficient to build an accurate model, because the high amount of parameters studied constrained their diversity in values. Nevertheless, the conclusions of the present work place us in a good starting point to follow the work, focusing on the parameters that are really significant.

The next step would be to develop a model that allows us to predict the quality as a function of σ and quantizer, starting from the knowledge gathered in this experiment. This model would be useful to select concrete values for these parameters, depending on the context. An example application may be to compute a good σ value for a low bitrate encoder.

Acknowledgements

The research leading to the presented results was partially supported by EU grants FP7 INFSO-ICT-248495. We would like to thank Juan Carlos Silva for the SciFi model and Peter Particle and Maya2OSG project for the Dragon model. We would also like to thank all volunteers that helped in the experiment, from syntheractive, CITIC research center, CINFO, R, Faculty of Computer Science and Faculty of Communication Sciences of University of A Coruña.

This paper was supported by the grant projects of the Czech Science Foundation no. 102/10/1320 *"Research and modeling of advanced methods of image quality evaluation (DEIMOS)"*, Ministry of Education, Youth and sports (MEYS) no. CZ.1.07/2.3.00/20.0007 *"Wireless Communication Team (WICOMT)"*, financed from the operational program Education for competitiveness, national project no. LD12005 *"Quality of Experience aspects of broadcast and broadband multimedia services (QUALEXAM)"* and by the internal grant of BUT project *FEKT-S-11-12 (MOBYS)*. The described research was also partly performed in laboratories supported by the SIX project; no. CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation.

Authors also wish to thank the anonymous reviewers for their useful comments that helped to improve the paper.

References

[1] ISO/IEC 13818-2. *Generic Coding of Moving Pictures and Associated Audio Information*. International Standard: ISO/IEC, 1995.

[2] WIEGAND, T., SULLIVAN, G. J., BJONTEGAARD, G., LUTHRA, A. Overview of the H.264/AVC video voding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 2003, vol. 13, no. 7, p. 560 - 576.

[3] SU, H., WEN, M., REN, J., WU, N., CHAI, J., ZHANG, C. Y. High-efficient parallel CAVLC encoders on heterogeneous multicore architectures. *Radioengineering*, 2012, vol. 21, no. 1, p. 46 - 55.

[4] MEENDERICK, C., AZEVEDO, A., ALVAREZ, M., JUURLINK, B., RAMIREZ, A. Parallel scalability of H.264. In *Proceedings of the 1st Workshop on Programmability Issues for Multi-Core Computers (MULTIPROG2008)*. Goteborg (Sweden), 2008, p. 1 - 12.

[5] REN, J., HE, Y., WU, W., WEN, M., WU, N., ZHANG, C. Software parallel CAVLC encoder based on stream processing. In *Proceedings of the 7th Workshop on Embedded Systems for Real-Time Multimedia (ESTIMedia2009)*. Grenoble (France), 2009, p. 126 - 133.

[6] MONTERO, P., GULIAS, V. M., TAIBO, J., RIVAS, S. Optimising lossless stages in a GPU-based MPEG encoder. *Multimedia Tools and Applications*, 2012, vol. 57, no. 2, p. 1 - 26.

[7] CHEN, T.-C., HUANG, Y.-W., TSAI, C.-Y., HSIEH, B.-Y., CHEN, L.-G. Dual-block-pipelined VLSI architecture of entropy coding for H.264/AVC baseline profile. In *Proceedings of the 15th International Symposium on VLSI Design, Automation and Test (2005 VLSI-TSA)*. Taiwan (China), 2005, p. 271 - 274.

[8] PRAKASH, V. A. M, GURUMURTHY, K. S. VLSI architecture for low power variable length encoding and decoding for image processing applications. *International Journal of Advances in Engineering and Technology*, 2012, vol. 2, no. 1, p. 105 - 120.

[9] LOHSCHELLER, H. A subjectively adapted image communication system. *IEEE Transactions on Communications*, 1984, vol. 32, no. 12, p. 1316 - 1322.

[10] CHANG, L.-W., WANG, CH.-Y.W., LEE, S.-M. Designing JPEG quantization tables based on human visual system. *Proceedings of the 5th International Conference on image Processing (ICIP99)*. Kobe (Japan), 1999, vol. 2, p. 376 - 380.

[11] MANNOS, J., SAKRISON, D. The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory*, 1974, vol. 20, no. 4, pp. 525 - 536.

[12] PITREY, Y., BARKOWSKY, M., PEPION, R., LE CALLET, P., HLAVACS, H. Influence of the source content and encoding configuration on the perceived quality for Scalable Video Coding. *Proceedings of the SPIE*, 2012, vol. 8291, p. 1 - 8.

[13] OU, Y.-F., ZENG, H., WANG, Y. Perceptual quality of video with quantization variation: A subjective study and analytical modeling. In *IEEE International Conference on Image Processing (ICIP2012)*. Orlando (FL, USA), 2011, p. 1 - 4.

[14] MONTERO, P., TAIBO, J., GULIAS, V. M., RIVAS, S. Parallel zigzag scanning and Huffman coding for a GPU-based MPEG-2 encoder. In *Proceedings of the IEEE International Symposium on Multimedia*. Dana Point (CA, USA), 2011, p. 97 - 104.

[15] REITER, U., KORHONEN, J. Comparing apples and oranges: subjective quality assessment of streamed video with different types of distortion. *International Workshop on Quality of Multimedia Experience (QoMEx)*. San Diego (CA, USA), 2009, p. 127 - 132.

[16] HUYNH-THU, I., GARCIA, M.-N., SPERANZA, F., CORRIVEAU, P., RAAKE, A. Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting*, 2011, vol. 57, no. 1, p. 1 - 14.

[17] WINKLER, S., MOHANDAS, P. The evolution of video quality measurement: from PSNR to hybrid metrics. *IEEE Transactions on Broadcasting*, 2008, vol. 54, no. 3, p. 46 - 51.

[18] SPERANZA, F., POULIN, F., RENAUD, R., CARON, M., DUPRAS, J. Objective and subjective quality assessment with expert and non-expert viewers. In *Proceedings of the 2nd International Workshop on Quality of Multimedia Experience (QoMEX)*. Trondheim (Norway), 2010, p. 46 - 51.

[19] ITU-R Recommendation BT.500-13. *Methodology for the Subjective Assessment of the Quality of Television Pictures*. Geneva (Switzerland): ITU, 2012.

[20] ITU-T Recommendation P.910. *Subjective Video Quality Assessment Methods for Multimedia Applications*. Geneva (Switzerland): ITU, 2008.

[21] SPERANZA, F., MARTIN, T., RENAUD, R. Subjective quality assessment and the effect of context in expert and non-expert viewers. In *Proceedings SPIE Image Quality and System Performance*. San Jose (USA), 2010, vol. 5294, p. 201 - 210.

[22] PINSON, M., WOLF, S. Comparing subjective video quality testing methodologies. In *Proceedings SPIE Visual Communications and Image Processing*. Lugano (Switzerland), 2003, vol. 5150, p. 573 - 582.

[23] SLANINA, M., KRATOCHVIL, T., POLAK, L., RICNY, V. Temporal aspects of scoring in the user based quality evaluation of HD video. In *Proceedings of the 34th Conference on Telecommunications and Signal Processing (TSP2011)*. Budapest (Hungary), 2011, p. 598 - 601.

[24] SLANINA, M., KRATOCHVIL, T., POLAK, L., RICNY, V. Analysis of temporal effects in quality assessment of high definition video. *Radioengineering*, 2012, vol. 21, no. 1, p. 63 - 69.

[25] THAKOLSRI, S., KELLERER, W., STEINBACH, E. QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation. In *IEEE International Conference on Communications (ICC)*. Kyoto (Japan), 2011, p. 1 - 6.

[26] PINSON, M. H., WOLF, S., CERMAK, G. HDTV subjective quality of H.264 vs. MPEG-2, with and without packet loss. *IEEE Transactions on Broadcasting*, 2010, vol. 56, no. 1, p. 86 - 91.

[27] SLANINA, M., KRATOCHVIL, T., BOLECEK, L, RICNY, V., KALLER, O., POLAK, L. Testing QoE in different 3D HDTV technologies. *Radioengineering*, 2012, vol. 21, no. 1, p. 445 - 454.

[28] *Living Lab Galicia.* [Online]. Available at: http://www.cinfo.es/?p=112&lang=en.

[29] SAKAMOTO, K., AOYAMA, S., ASAHARA, S., YAMASHITA, K., OKADA, A. Evaluation of viewing distance vs. TV size on visual fatigue in a home viewing environment. In *Digest of Technical Papers, International Conference on Consumer Electronics (ICCE09)*. Las Vegas (USA), 2009, p. 1 - 2.

[30] KLIMA, M. et al. DEIMOS - an open source image database. *Radioengineering*, 2011, vol. 20, no. 4, p. 1016 - 1023.

[31] COX, D. R. *The Collected Works of John W. Tukey: Factorial and Anova, Volume VII (Statistics/probability series)*. Chapman & Hall/CRC, 1992, 350 pages.

[32] SHAPIRO, S. S., WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, 1965, vol. 52, no. 3/4, p. 591 - 611.

[33] BROWN, M. B., FORSYTHE, A. B. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 1974, vol. 69, no. 346, p. 364 - 367.

[34] KRUSKAL, W. H., WALLIS, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 1952, vol. 69, no. 260, p. 583 - 621.

About Authors ...

Pablo MONTERO was born in Ferrol (Spain). He received B. Eng degree in computer engineering from Universidade da Coruña, Spain, in 2009. He is currently a Ph.D candidate in this University. His research work is focused on parallel algorithms for video encoding, working at CITIC IT research center at A Coruña.

Ladislav POLÁK was born in Štúrovo (Slovakia), in 1984. He received the M.Sc. degree in 2009 in Electronics and Communications program from the Brno University of Technology. He is currently Ph.D. student at the Department of Radio Electronics, Brno University of Technology. His research interests include digital television and audio broadcasting, wireless communication and mobile systems, video and multimedia transmission, including video image quality evaluation. He has been an IEEE student member since 2010.

Javier TAIBO was born in A Coruña (Spain), in 1973. He received the B.S. Degree in Computer Science in 1998 and later the Ph.D. in Computer Science in 2010 from University of A Coruña. His research interests are related to many fields of multimedia and computer graphics, oriented towards interactive applications and real-time rendering/encoding. He is a full-time professor in the University of A Coruña, teaching several subjects related to 3D computer animation.

Tomáš KRATOCHVÍL was born in Brno (Czech Republic), in 1976. He received the M.Sc. degree in 1999, Ph.D. degree in 2006 and Assoc. Prof. position in 2009, all in Electronics and Communications program from the Brno University of Technology. He is currently an associated professor at the Department of Radio Electronics, Brno University of Technology. His research interests include digital television and audio broadcasting, its standardization and video and multimedia transmission including video image quality evaluation. He has been an IEEE member since 2001.