Subjective Impression of Variations in Layer Encoded Videos

Michael Zink, Oliver Künzel, Jens Schmitt, Ralf Steinmetz
KOM Multimedia Communications
Darmstadt University of Technology
Merckstrasse 25, D-64283 Darmstadt, Germany
{zink, okuenzel, schmitt, steinmetz}@kom.tu-darmstadt.de

ABSTRACT

Layer encoded video is an elegant way to allow adaptive transmissions in the face of varying network conditions as well as it supports heterogeneity in networks and clients. As a drawback quality degradation can occur, caused by variations in the amount of transmitted layers. Recent work on reducing these variations makes assumptions about the perceived quality of those videos. The main goal of this paper respectively its motivation is to investigate the validity of these assumptions by subjective assessment. However, the paper is also an attempt to investigate fundamental issues for the human perception of layer encoded video with time-varying quality characteristics. For this purpose, we built a test environment for the subjective assessment of layer encoded video and conducted an empirical experiment in which 66 test candidates took part. The results of this subjective assessment are presented and discussed. To a large degree we were able to validate existing (unproven) assumptions about quality degradation caused by variations in layer encoded videos, however there were also some interesting, at first sight counterintuitive findings from our experiment.

Keywords

Empirical experiment, layer encoded video, human perception, video quality variations.

1. INTRODUCTION

1.1 Motivation

In the area of video streaming layer encoded video is an elegant way to overcome the inelastic characteristics of traditional video encoding formats like MPEG-1 or H.261. Layer encoded video is particularly useful in today's Internet where a lack of Quality of Service (QoS) mechanisms might make an adaptation to existing network conditions necessary. In addition, it bears the capability to support a large variety of clients while only a single file[1] has to be stored at a video server for each video object. The drawback of adaptive transmissions is the introduction of variations in the amount of transmitted layers during a streaming session. These variations affect the end-user's perceived quality and thus the acceptance of a service that is based on such technology.

Recent work that has focused on reducing those layer variations, either by employing intelligent buffering techniques at the client [2, 3, 4] or proxy caches [5, 6, 7] in the distribution network, made various assumptions about the perceived quality of videos with time-varying number of layers. To the best of our knowledge, these assumptions have not been verified by subjective assessment so far.

The lack of in-depth analysis about quality metrics for variations in layer encoded videos led us to conduct an empirical experiment based on subjective assessment to obtain results that can be used in classifying the perceived quality of such videos.

1.2 What is the Relation between Objective and Subjective Quality?

The goal of this research work is to investigate if general assumptions made about the quality metrics of variations in layer encoded videos can be verified by subjective assessment. We use the following example to explain our intention in more detail: A layer encoded video that is transmitted adaptively[2] to the client might have layer variations as shown in Figure 1. In Section 2.1 several quality metrics that allow the determination of the video's quality are presented. At first, we discuss the basics of these quality metrics. The most straightforward quality metric would be the total sum of all received segments (see Figure 1). However, common assumptions on the quality of a layer encoded video are that the quality is not only influenced by the total sum of received segments but also by the frequency of layer variations and the amplitude of those variations [3, 5, 7]. As shown in Figure 1 the amplitude specifies the height of a layer variation while the frequency determines the amount of layer variations.

All quality metrics we are aware of are based on these assumptions. Verifying all possible scenarios that are covered by those assumptions with an experiment based on subjective assessment is hard to achieve. Therefore, we decided to focus on basic scenarios that have the potential to answer the most fundamental questions, e.g., are the

[1] In contrast to the dynamic stream switching [1] approach where for each quality level one specific video file is required.

[2] Adaptively in this case means that the amount of layers transmitted to the client is based on some feedback from the network or the client, e.g., congestion control information.

sequences on the left in Figure 2 ((a1) and (b1)) more annoying than sequences on the right ((a2) and (b2)) for an end-user who views a corresponding video sequence. In this example, the first scenario ((a1) and (a2)) is focussed on the influence of the amplitude and the second ((b1) and (b2)) on the frequency of layer variations.

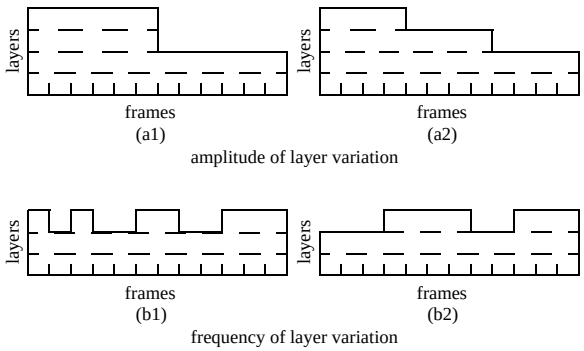


Figure 2. Quality criteria [3]

1.3 Outline

The paper is structured as follows. Section 2 reviews previous work on retransmission scheduling for layer encoded video and subjective assessment of video quality. The test environment and the subjective test method used for the experiment are described and discussed in Section 3. The details of the experimental setup are given in Section 4 and in Section 5 the results of the experiment are presented and discussed. Section 6 summarizes the major conclusions that can be drawn from the experiment.

2. RELATED WORK

The related work section is split in two parts since our work is influenced by the two research areas briefly surveyed in the following.

2.1 Retransmission Scheduling

The work presented in this paper has been motivated by our own work on quality improvement for layer encoded videos. During our investigation of favorable retransmission scheduling algorithms which are supposed to improve the quality of layer encoded videos stored on a cache [7], we realized that in related work the quality metrics for layer encoded videos are based on somewhat speculative assumptions only. To the best of our knowledge none of these assumptions is based on a subjective assessment.

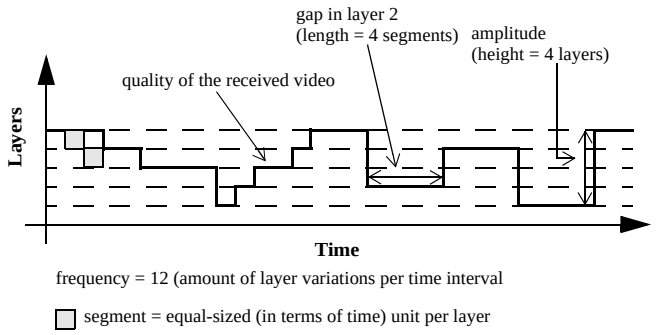


Figure 1. Quality of a layer encoded video at the client

In [3], Nelakuditi et al. state that a good metric should capture the amount of detail per frame as well as its uniformity across frames. I.e., if we compare the sequences of layers in a video shown in Figure 2 the quality of (a2) would be better than that of (a1) which is also valid for (b2) and (b1), according to their assumption. Their quality metric is based on the principle of giving a higher weight to lower layers and to longer runs of continuous frames in a layer.

The metric presented by the work of Rejaie et al. [5] is almost identical to the one advocated for in [3]. *Completeness* and *continuity* are the 2 parameters that are incorporated in this quality metric. *Completeness* of a layer is defined as the ratio of the layer size transmitted to its original (complete) size. E.g. the ratio of layer 2 in sequence (a1) in *Figure 2* would be 1 while the ratio for layer 3 would be 0.5. *Continuity* is the metric that covers the 'gaps' in a layer. It is defined as the average number of segments between two consecutive layer breaks (i.e., gaps). In contrast to the other metrics presented here, this metric is a per-layer metric.

In our previous work [7] we also made assumptions about the quality metrics for layer encoded videos. Similar to [3] we postulated that this metric should be based on a) the frequency of variations and b) the amplitude of variations.

2.2 Video Quality

There has been a substantial amount of research on methodologies for subjective assessment of video quality, e.g., [8] and [9], which contributed to form an ITU Recommendation on this issue [10]. This standard has been used as a basis for subjective assessment of encoders for digital video formats, in particular for MPEG-2 [11, 12] and MPEG-4 [13] but also on other standards like H.263+ [14]. The focus of interest for all these subjective assessment experiments was the quality of different coding and compression mechanisms. Our work, in contrast, is concerned with the quality degradation caused by variations in layer encoded videos. Like us, [15] is also concerned with layer encoded video and presents the results of an empirical evaluation of 4 hierarchical video encoding schemes. This is orthogonal to our work since the focus of their investigation is on the comparison between the different layered coding schemes and not on the human perception of layer variations.

In [16], a subjective quality assessment has been carried out in which the influence of the frame rate on the perceived quality is investigated. In contrast to our work elasticity in

the stream was achieved by frame rate variation and not by applying a layer encoded video format.

Effects of bit errors on the quality of MPEG-4 video were explored in [17] by subjective viewing measurements, but effects caused by layer variations were not examined.

Chen presents an investigation on an IP-based video conference system [18]. The focus in this work is mainly auditorium parameters like display size and viewing angle. A layer encoded video format is not used in this investigation.

Probably closest to our work, Lavington et al. [19] used an H.263+ two layer video format in their trial. In comparison to our approach, they were rather interested in the quality assessment of longer sequences (e.g., 25 min.). Instead of using identical pregenerated sequences that were presented to the test candidates, videos were streamed via an IP network to the clients and the quality was influenced in a fairly uncontrolled way by competing data originating from a traffic generator. The very specific goal of this work was to examine if reserving some of the network's bandwidth for either the base or the enhancement layer improves the perceived quality of the video, while we are rather interested on the influence of variations in layer encoded videos and try to verify some of the basic assumption made about the perceived quality in a subjective assessment experiment. Furthermore, we try to conduct this experiment in a much more controlled environment in order to achieve more significant and easier to interpret results.

3. TEST ENVIRONMENT

In this section, we first present the layer encoded video format used for the experiment, describe how we generated the test sequences, explain why we decided to use stimulus-comparison as the assessment method, and shortly present our test application.

3.1 Layer Encoded Video Format - SPEG

SPEG (Scalable MPEG) [20] is a simple modification to MPEG-1 which introduces scalability. In addition to the possibility of dropping complete frames (temporal scalability), which is already supported by MPEG-1 video, SNR scalability is introduced through layered quantization of DCT data [20]. The extension to MPEG-1 was made for two reasons. First, there are no freely available implementations of layered extensions for existing video standards (MPEG-2, MPEG-4), second, the granularity of scalability is improved by SPEG combining temporal and SNR scalability. As shown in Figure 3 a priority (p_0 - p_{11}) can be mapped to each layer. The QoS Mapper (see Figure 4, which depicts the SPEG pipeline and its components) uses the priority information to determine which layers are dropped and which are forwarded to the Net Streamer.

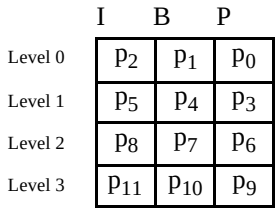


Figure 3. SPEG layer model

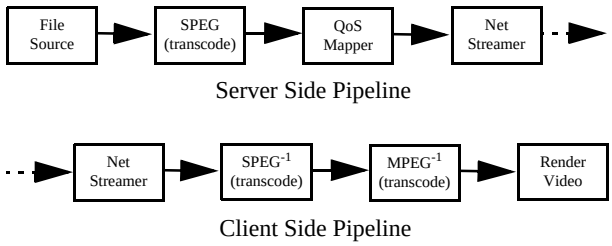


Figure 4. Pipeline for SPEG [21]

Our decision to use SPEG as a layer encoded video format is based on several reasons. SPEG is designed for a QoS-adaptive video-on-demand (VoD) approach, i.e., the data rate streamed to the client should be controlled by feedback from the network (e.g., congestion control information). In addition, the developers of SPEG also implemented a join function that re-transcodes SPEG into MPEG-1 [21] and therefore allows the use of standard MPEG-1 players, e.g., the Windows Media Player. We were not able to use scalable video encoders available as products (e.g., [22, 23]) because videos created by those can only be streamed to the corresponding clients which do neither allow the storage of the received data on a disk nor the creation of scheduled quality variations.

3.2 Test Generation - Full Control

Since our test sequences must be created in a deterministic manner, we slightly modified the SPEG pipeline. The most important difference is, that in our case data belonging to a certain layer must be dropped intentionally and not by an unpredictable feedback from the network or the client. This modification was necessary, since identical sequences must be presented to the test candidates in the kind of subjective assessment method that is used in our experiment. Therefore, we modified the QoS Mapper in a way that layers are dropped at certain points in time specified by manually created input data. We also added a second output path to the MPEG^{-1} module that allows us to write the resulting MPEG-1 data in a file.

3.3 Measurement Method -

Stimulus Comparison

The subjective assessment method is widely accepted for determining the perceived quality of images and videos. Research that was performed under the ITU-R lead to the development of a standard for such test methods [10]. The standard defines basically five different test methods double-stimulus impairment scale (DSIS), double-stimulus continuous quality-scale (DSCQS), single stimulus quality evaluation (SSCQE), simultaneous double stimulus for continuous evaluation (SDSCE), and stimulus-comparison (SC), respectively.

Since it was our goal to investigate the basic assumptions about the quality of layer encoded video, SSCQE and SDSCE are not the appropriate assessment method because comparisons between two videos are only possible on an identical time segment and not between certain intervals of the same video. In addition, SSCQE and SDSCE were designed to assess the quality of an encoder (e.g., MPEG-1) itself.

Two test methods which better suit the kind of investigations we want to perform are DSCQS and DSIS. Compared to SSCQE and SDSCE they allow to asses the quality of a codec in relation to data losses [8] and therefore, are more suitable if the impairment caused by the transmission path is investigated.

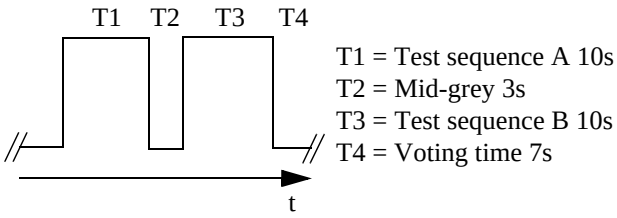


Figure 5. Presentation structure of test material

The SC method differs from DSCQS and DSIS in a way that two test sequences with unequal qualities are shown (see Figure 5) and the test candidates can vote on a scale as shown in Table 1. Comparing two impaired videos directly with each other is our primary goal. Since this is represented best by the SC method we decided to use this method in our test.

Additionally, preliminary tests have shown us that test candidates with experience in watching videos on a computer are less sensitive to impairment. I.e., they recognize the impairment but do not judge it as annoying as candidates who are unexperienced. This effect is dampened since only impaired sequences have to be compared with each other in a single test that is based on the SC method. Our preliminary tests with the DSIS method, where always the original sequence and an impaired sequence are compared, delivered results with less significance compared to tests performed with the SC method

Table 1: Comparison scale

Value	Compare
-3	much worse
-2	worse
-1	slightly worse
0	the same
1	slightly better
2	better
3	much better

3.4 Test Application - Enforcing Time Constraints

We created a small application[3] (see Figure 6) that allows an automated execution of the tests. Since we had to use a computer to present the videos anyway, we decided to let the candidates perform their voting also on the computer. Using this application has the advantage that we can easily enforce the time constraints demanded by the measurement method, because we allow voting only during a certain time interval. As a convenient side effect, the voting data is available in a machine readable format.

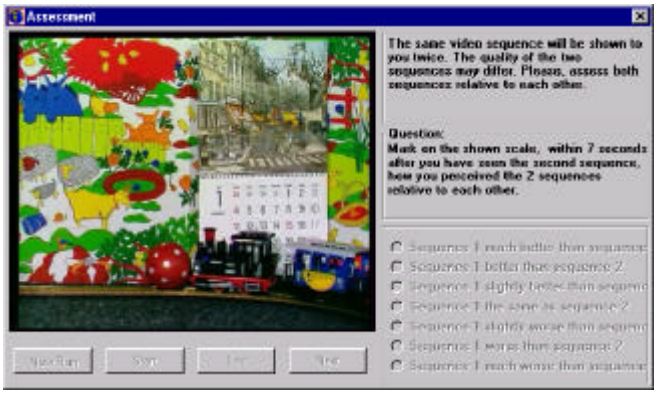


Figure 6. Application for experiment

4. EXPERIMENT

4.1 Scenario

Since quality metrics for layer encoded video are very general, we have to focus on some basic test cases in order to keep the amount of tests that should be performed in the experiment feasible. We decided to investigate isolated effects, one-by-one at a time, which on one hand keeps the

[3] A downloadable version of the test can be found at http://www.kom.e-technik.tu-darmstadt.de/video-assessment/

size of a test session reasonable and on the other hand still allows to draw conclusions for the general assumptions, as discussed above. That means we are rather interested in observing the quality ranking for isolated effects like frequency variations (as shown in sequences (b1) and (b2) in Figure 2) than for combined effects (as shown in Figure 1). This bears also the advantage that standardized test methods [10], which limit the sequence length to several seconds, can be applied. All patterns that were used for the experiment are shown in Figure 8.

4.2 Candidates

The experiment was performed with 66 test candidates (45 males and 21 females), between the age of 14 and 64. 55 of them had experiences with watching videos on a computer.

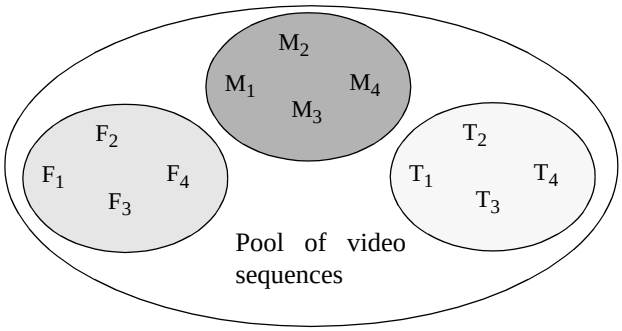
4.3 Procedure

Each candidate had to perform 15 different assessments, of which each single test lasted for 33 seconds. All 15 tests were executed according to the SC assessment method. The complete test session per candidate lasted for about 15 minutes[4], on average. We have chosen three video sequences for this experiment, that have been frequently used for subjective assessment [24]. The order of the 15 video sequences was changed randomly from candidate to candidate as proposed in the ITU-R B.500-10 standard [10] (see also Figure 7). After some initial questions (age, gender, profession) 3 assessments were executed as a warm-up phase. This should avoid that the test candidates are distracted by the content of the video sequences as reported by Aldridge et al. [11]. In order to avoid that two consecutive video sequences (e.g., F_2 is following F_1 immediately) have the same content we defined a pattern for the chronological order of the test sessions, as shown in Figure 7. F_x can be any video sequence from the F pool of sequences that has not been used in this specific test session, so far. Thus, a complete test session for a candidate could have a chronological order as shown for Figure 7.

4.4 Layer Patterns

Figure 8 shows the layer patterns of each single sequence that was used in the experiment, except for the first 3 warm-up tests where the comparison is performed between the first sequence that consists of 4 layers and the second that consists of only one layer. Each of the 3 groups shows the patterns that were used with one type of content. Comparisons were always performed between patterns that are shown in a row (e.g., (a1) and (a2)). As already mentioned in Section 1.2 it was our goal to examine fundamental assumptions about the influence of layer

[4] Only watching the sequences and voting took less time, but the candidates had as much time as they wanted to read the questions and possible answers for each test ahead of each test.



Pattern	I_1	I_2	I_3	F_x	M_x	T_x	F_x	M_x	T_x	F_x	M_x	T_x	F_x	M_x	T_x
Sequence 1	I_1	I_2	I_3	F_3	M_1	T_4	F_1	M_2	T_1	F_4	M_3	T_2	F_2	M_4	T_3
Sequence 2	I_1	I_2	I_3	F_3	M_1	T_4	F_1	M_2	T_1	F_4	M_3	T_2	F_2	M_4	T_3

F = Farm
M = Mobile & Calendar
T = Table Tennis

Figure 7. Random generation of test sequence order

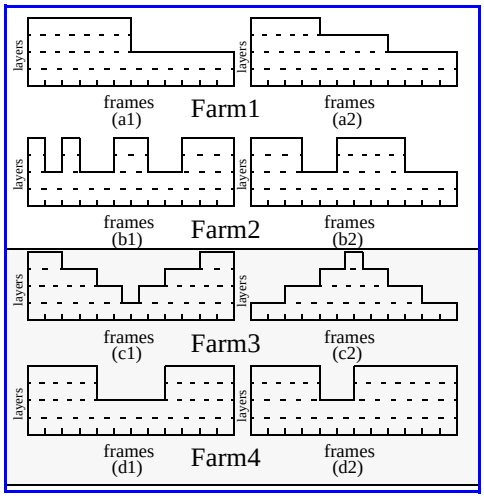
changes on perceived quality. This is also reflected by the kind of patterns we decided to use in the experiment. It must be mentioned that the single layers are not equal in size (contrary to the presentation in Section 8). The size of the n^{th} layer is rather determined by the following expression: $s_n = 2s_{n-1}$. Thus, segments of different layers have different sizes. Preliminary experiments have shown that equal layer sizes are not appropriate to make layer changes perceivable. Since there exist layered schemes that produce layers with sizes similar to ours [25, 26], we regard this a realistic assumption.

In the experiment, we differentiate between two groups of tests, i.e., one group in which the amount of segments used by a pair of sequences is equal and one in which the amount differs (the latter has a shaded background in Figure 8). We made this distinction because we are mainly interested in how the result of this experiment could be used to improve the retransmission scheduling technique (see Section 2.1) where it is necessary to compare the influence of additional segments that is added on different locations in a sequence.

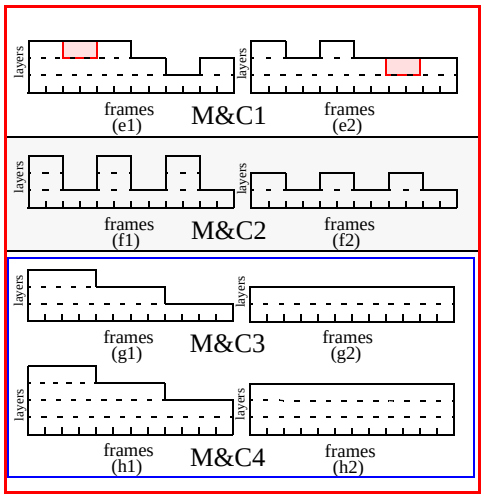
Since segments from different layers are not equal in size, the amount of data for the compared sequences differs. However, somewhat surprisingly, as we discuss in Section 5.3, a larger amount of data does not necessarily lead to a better perceived quality. Additional tests with different quantities of segments in between a pair were chosen to answer additional questions and make the experiment more consistent as we show in Section 5.2.

5. RESULTS

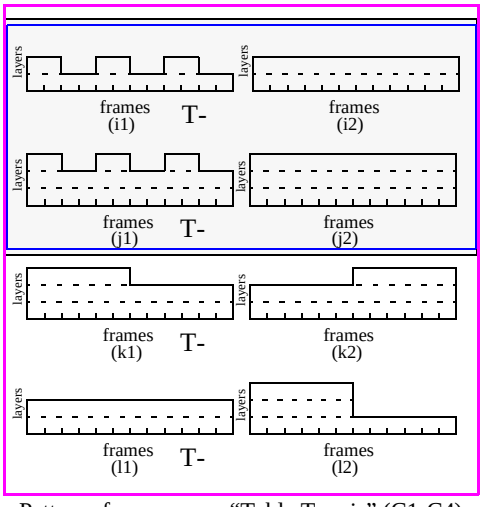
In this section, we present the results of the experiment described in Section 4. Since we analyze the gathered data statistically it must clearly be mentioned that the presented results cannot prove an assumption but only make it less or



Patterns for Sequence “Farm” (F1-F4)



Patterns for sequence “Mobile & Calendar” (M1-M4)



Patterns for sequence “Table Tennis” (C1-C4)

Figure 8. Segments that were compared in the experiment

more likely based on the gathered data. The overall results of all experiments are summarized in Figure 9 and are discussed in the following subsections.

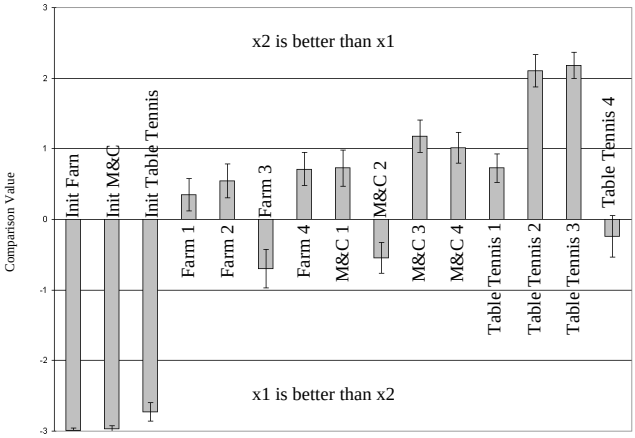


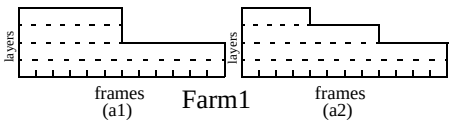
Figure 9. Average and 95% confidence interval for the different tests of the experiment

5.1 Same Amount of Segments

In this section, we discuss the results for the assessments of tests in which the total sum of segments is equal. That means the space covered by the pattern of both sequences is identical.

5.1.1 Farm1: Amplitude

In this assessment the stepwise decrease was rated slightly better than one single but higher decrease. The result shows a tendency that the assumptions that were made about the amplitude of a layer change (as described in Section 2.1) are correct.



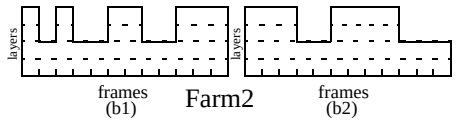
Average:0.35 Conf.-Interval Size:0.23

Figure 10. Farm1

5.1.2 Farm2: Frequency

The result of this test has an even higher likelihood that the second sequence has a better perceived quality than it is the case for *Farm1*. It tends to confirm the assumption that the frequency of layer changes influences the perceived quality,

since, on average, test candidates ranked the quality of the sequence with lesser layer changes better.

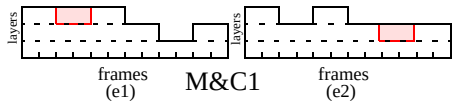


Average:0.55 Conf.-Interval Size:0.24

Figure 11. Farm2

5.1.3 M&C1: Closing the gap

This test should try to answer the question, if it would be better to close a gap in a layer on a higher or lower level. The majority of the test candidates decided that filling the gap on a lower level results in a better quality than otherwise. This result tends to affirm our assumptions made for retransmission scheduling in [7].

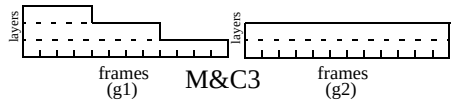


Average:0.73 Conf.-Interval Size:0.25

Figure 12. M&C1

5.1.4 M&C3: Constancy

Even more significant than in the preceding tests, the candidates favored the sequence with no layer changes as the one with the better quality. One may judge this a trivial and unnecessary test, but from our point of view the result is not that obvious, since (g1) starts with a higher amount of layers. The outcome of this test implies that it might be better, in terms of perceived quality, to transmit less but a constant amount of layers.

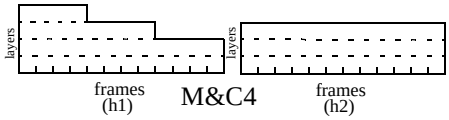


Average:1.18 Conf.-Interval Size:0.23

Figure 13. M&C3

5.1.5 M&C4: Constancy at a higher level

This test was to examine if an increase of the overall level (in this case by comparison to Section 5.1.4) has an influence on the perceived quality. Comparing the results of both tests (*M&C3* and *M&C4*) shows no significant change in the test candidates' assessment. 66% of the test candidates judge the second sequences ((g2) and (h2)) better (values 1-3 in Table 1) in both cases which makes it likely that the overall level has no influence on the perceived quality.

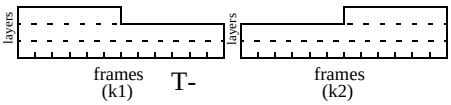


Average:1.02 Conf.-Interval Size:0.22

Figure 14. M&C4

5.1.6 Tennis3: All is well that ends well

The result of this test shows the tendency that increasing the amount of layers in the end leads to a higher perceived quality. The result is remarkably strong (the highest bias of all tests). Future tests, that will be of longer duration and executed in a different order (first (k2) than (k1)), will show how the memory-effect [11] of the candidates influenced this test.

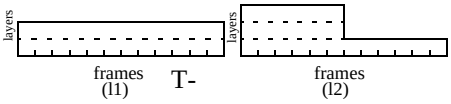


Average:2.18 Conf.-Interval Size:0.19

Figure 15. Tennis3

5.1.7 Tennis4: The exception proves the rule

This test is the only one out of the 12 tests in which the 95% confidence interval covers both areas (better, worth) of the judgement scale. If we regard the average only, the result is a little bit surprising since it contradicts the results from Section 5.1.1 and Section 5.1.4, respectively. At this stage of the investigation, we can only assume that also the content might have an influence on the perceived quality. But to gain more insight in this phenomenon further experiments are necessary.



Average:-0.24 Conf.-Interval Size:0.29

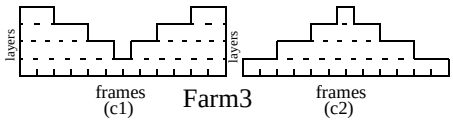
Figure 16. Tennis4

5.2 Different amount of Segments

In the following 5 tests the total amount of segments per sequence differs. All 5 tests have in common that the perceived quality of the sequence consisting of a pattern that covers a larger number of segments were ranked better.
This is obvious, but it makes the overall result more consistent, because test candidates mostly realized this quality difference.

5.2.1 Farm3: Decrease vs. increase

Starting with a higher amount of layers, decreasing the amount of layers, and increasing the amount of layers in the end again seems to provide a better perceivable quality than starting with a low amount of layers, increasing this amount of layers, and going back to a low amount of layers at the end of the sequence. This might be caused by the fact that test candidates are very concentrated in the beginning and the end of the sequence and that, in the first case details become clear right in the beginning of the sequence.

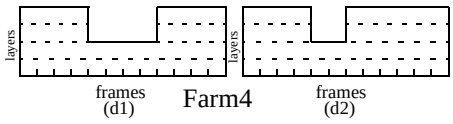


Average:-0.70 Conf.-Interval Size:0.27

Figure 17. Farm3

5.2.2 Farm4: Keep the gap small

In this test, it was our goal to investigate how the size of a gap may influence the perceived quality. The majority of test candidates (37 out of 66) judged the quality of the sequence with a smaller gap slightly better (Only 5 out of 66 judged the first sequence better). This indicates that filling a gap partly can be beneficial.

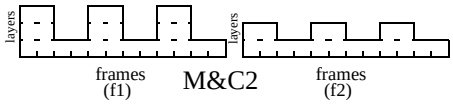


Average:0.71 Conf.-Interval Size:0.23

Figure 18. Farm4

5.2.3 M&C2: Increasing the amplitude

The effect of the amplitude height should be investigated in this test. The result shows that, in contrast to existing assumptions (see Section 2.1), an increased amplitude can lead to a better perceived quality.

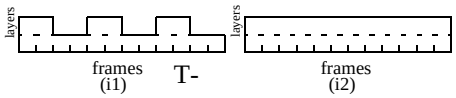


Average:-0.55 Conf.-Interval Size:0.22

Figure 19. M&C2

5.2.4 Tennis1: Closing all gaps

This test is contrary to *M&C2* where the additional segments are used to close the existing gaps instead of increasing the amplitude of already better parts of the sequence. This strategy decreases the frequency of layer changes. Test candidates, on average, judged the sequence without layer changes better. The result of this test reaffirms the tendency that was already noticed in Section 5.1.2, that the perceived quality is influenced by the frequency of layer changes. If we carefully compare the results of *M&C2* and *Tennis1*, a tendency towards filling the gaps and thus decreasing the frequency instead of increasing the amount of already increased parts of the sequence is recognizable. Definitely, further investigations are necessary to confirm this tendency, because, here, the results of tests with different contents are compared and we have not investigated the influence of the content on the perceived quality, so far.

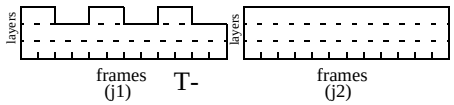


Average:0.73 Conf.-Interval Size:0.20

Figure 20. Tennis1

5.2.5 Tennis2: Closing all gaps at a higher level

In comparison to *Tennis1*, here, we were interested in how an overall increase of the layers (in this case by one layer) would influence the test candidates judgement. Again the sequence with no layer changes is judged better, even with a higher significance than for *Tennis1*. This might be caused by the fact that the amount of layer is higher in general in *Tennis2*.



Average:2.11 Conf.-Interval Size:0.23

Figure 21. Tennis1

5.3 Sequence Size and Quality

As already mentioned in Section 4.4, segments of different layers are not equal in size. Hence, the data size for patterns with an equal amount of segments might not be identical. Here we give an example that shows that sequences consisting of a higher amount of data are not necessarily judged better concerning the perceived quality of the sequence. If we compare, e.g., the tests *M&C1, M&C3,* and *M&C4,* respectively, we recognize that the storage size of the first sequence is always larger than the one of the second. This relation is shown in Figure 22. The results of our experiment show that the average judgement is in contrast to the storage size (Figure 8) where the second sequence has, according to the test candidates, a better perceived quality.

6. CONCLUSION

In this paper, we presented the results of an empirical experiment based on subjective assessment of variations in layer encoded video. A statistical analysis of the experiment

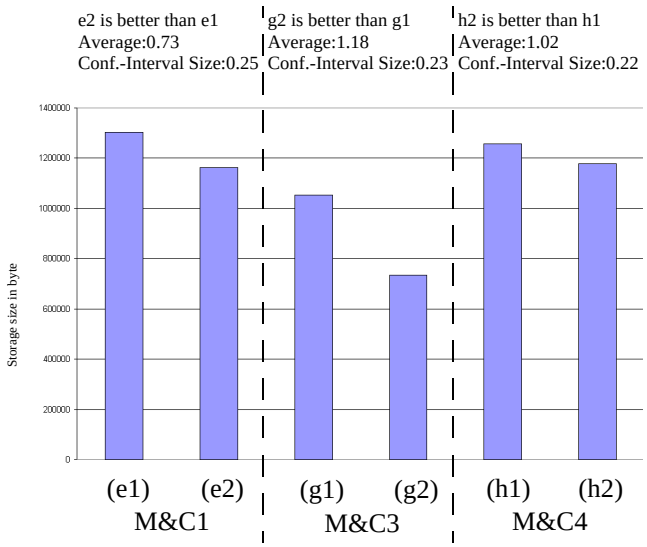


Figure 22. Storage size of sequences

mostly validates assumptions that were made in relation to layer variations and the perceived quality of a video:

- The frequency of variations should be kept as small as possible.
- If a variation can not be avoided the amplitude of the variation should be kept as small as possible.

One basic conclusion from the results in Section 5.2 is: adding information to a layered video increases its average quality. But, as we already assumed in our work on retransmission scheduling, adding information at different locations can have a substantial effect on the perceived quality. Assumptions we made for our heuristics in retransmission scheduling (as well as others' assumptions) could be substantiated by this investigation (see Section 5.1). That means, it is more likely that the perceived quality of a layer encoded video is improved if

- the lowest quality level is increased, and
- gaps in lower layers are filled.

The results from Section 5.3 should be used to refine the retransmission scheduling heuristics in relation to the size of each single layer. Therefore, the metric that represents the quality improvement must also take into account that it might be more expensive to retransmit a segment of layer n+1 than of layer n. Another interesting outcome of the experiment is the fact that a quality improvement may be achieved by retransmitting less data (Section 5.1.3), if a layered encoding scheme is used in which the layers are not of identical size. The obtained results can, in addition, be used to refine caching replacement policies that operate on a layer level [5] as well as layered multicast transmission schemes which try to offer heterogeneous services to different subscribers as, e.g., in the receiver-driven layered multicast RLM [27] scheme and its derivations.

The results of this investigation clearly strengthen the assumption that a differentiation between objective and subjective quality, in the case of variations in layer encoded video, must be made.

Nevertheless, it must be admitted that the presented work is only an initial investigation in the subjective impression of variations in layer encoded videos. In further work, we want to explore sequences with a longer duration (up to several minutes). In a next step, we will investigate if the shown sequences can be combined and if the subjective assessment is still consistent with the separated results. E.g., in this experiment sequences (e2) and (g2) were judged better than (e1) and (g1), will a sequence that combines (e2) and (g2) also be judged better than a sequence that combines (e1) and (g1)? We are also interested in how the content of a sequence influences the perceived quality.

7. ACKNOWLEDGMENTS

The authors would like to thank Rico Tunk for creating the test application, Charles "Buck" Krasic for his support on SPEG, the test candidates for taking the time to perform the assessment, and RTL Television for providing the video sequences.

8. REFERENCES

[1] J. Lu. Signal Processing for Internet Video Streaming: A Review. In *Proceedings of SPIE Image and Video Communications and Processing, San Jose, CA, USA*. SPIE - The International Society for Optical Engineering, January 2000.

[2] D. Saparilla and K. W. Ross. Optimal Streaming of Layered Video. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies 2000 (INFOCOM'00), Tel-Aviv, Israel*, pages 737–746, March 2000.

[3] S. Nelakuditi, R. R. Harinath, E. Kusmierek, and Z.-L. Zhang. Providing Smoother Quality Layered Video Stream. In *Proceedings of the 10th International Workshop on Network and Operating System Support for Digital Audio and Video, Raleigh, NC, USA*, June 2000.

[4] R. Rejaie, M. Handley, and D. Estrin. Quality Adaptation for Congestion Controlled Video Playback over the Internet. In *Proceedings of the ACM SIGCOMM '99 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication 1999, New York, NY, USA*, pages 189–200, August 1999.

[5] R. Rejaie, H. Yu, M. Handley, and D. Estrin. Multimedia Proxy Caching for Quality Adaptive Streaming Applications in the Internet. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies 2000 (INFOCOM'00), Tel-Aviv, Israel*, pages 980–989, March 2000.

[6] R. Rejaie and J. Kangasharju. Mocha: A Quality Adaptive Multimedia Proxy Cache for Internet Streaming. In *Proceedings of the 11th International Workshop on Network and Operating System Support for Digital Audio and Video, Port Jefferson, New York, USA*, pages 3–10, June 2001.

[7] M. Zink, J. Schmitt, and R. Steinmetz. Retransmission Scheduling in Layered Video Caches, April 2002. Accepted at ICC 2002, New York, New York, USA.

[8] T. Alpert and J.-P. Evain. Subjective quality evaluation - The SSCQE and DSCQE methodologies. EBU Technical Review, February 1997.

[9] E. Kohler, M. Handley, S. Floyd, and J. Padhye. Datagram Control Protocol (DCP). Internet Draft, November 2001. Work in Progress.

[10] ITU-R: Methodology for the Subjective Assessment of the Quality of Television Picture. International Standard, 2000. ITU-R BT.500-10.

[11] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson. Measurement of scene-dependent quality variations in digitally coded television pictures. *IEE Proceedings on Vision, Image and Signal Processing*, 142, 3:149–154, 1995.

[12] R. Aldridge, D. Hands, D. Pearson, and N. Lodge. Continuous quality assessment of digitally-coded television pictures. *IEE Proceedings on Vision, Image and Signal Processing*, 145, 2:116–123, 1998.

[13] F. Pereira and T. Alpert. MPEG-4 video subjective test procedures and results. *IEEE Transactions on Circuits and Systems for Video Technology*, 7, 1:32–51, 1997.

[14] M. Masry and S. Hemami. An analysis of subjective quality in low bit rate video. In *International Conference on Image Processing (ICIP), 2001, Thessaloniki, Greece*, pages 465–468. IEEE Computer Society Press, October 2001.

[15] C. Kuhmünch and C. Schremmer. Empirical Evaluation of Layered Video Coding Schemes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece*, pages 1013–1016, October 2001.

[16] T. Hayashi, S.Yamasaki, N. Morita, H. Aida, M. Takeichi, and N. Doi. Effects of IP packet loss and picture frame reduction on MPEG1 subjective quality. In *3rd Workshop on Multimedia Signal Processing, Copenhagen, Denmark*, pages 515–520. IEEE Computer Society Press, September 1999.

[17] S. Gringeri, R. Egorov, K. Shuaib, A. Lewis, and B. Basch. Robust compression and transmission of MPEG-4 video. In *Proceedings of the ACM Multimedia Conference 1999, Orlando, Florida, USA*, pages 113–120, October 1999.

[18] M. Chen. Design of a virtual auditorium. In *Proceedings of the ACM Multimedia Conference 2001, Ottawa, Canada*, pages 19–28, September 2001.

[19] S. Lavington, N. Dewhurst, and M. Ghanbari. The Performance of Layered Video over an IP Network. *Signal Processing: Image Communication, Elsevier Science*, 16, 8:785–794, 2001.

[20] C. Krasic and J. Walpole. Priority-Progress Streaming for Quality-Adaptive Multimedia. In *ACM Multimedia Doctoral Symposium, Ottawa, Canada*, October 2001.

[21] C. Krasic and J. Walpole. QoS Scalability for Streamed Media Delivery. Technical Report OGI CSE Technical Report CSE-99-011, Oregon Graduate Institute of Science & Technology, September 1999.

[22] Intel. Developers - What Intel Streaming Web Video Software Can Do For You, 2000. http://developer.intel.com/ial/swv/developer.htm.

[23] PacketVideo. Technical White Paper: PacketVideo Multimedia Technology Overview, 2001. http://www.packetvideo.com/pdf/pv_whitepaper.pdf.

[24] R. Neff and A. Zakhor. Matching Pursuit Video Coding–Part I: Dictionary Approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 12, 1:13–26, 2002.

[25] J. Hartung, A. Jacquin, J. Pawlyk, and K. Shipley. A Real-time Scalable Software Video Codec for Collaborative Applications over Packet Networks. In *Proceedings of the ACM Multimedia Conference 1998, Britol, UK*, pages 419–426, September 1998.

[26] L. Vicisano, L. Rizzo, and J. Crowcroft. TCP-like congestion control for layered multicast data transfer. In *Proceedings of the 17th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'98)*, pages 996–1003. IEEE Computer Society Press, March 1998.

[27] S. McCanne, M. Vetterli, and V. Jacobson. Receiver-driven layered multicast. In *Proceedings of ACM SIGCOMM'96*, Palo Alto, CA, August 1996.

Subjective Impression of Variations in Layer Encoded Videos

Michael Zink, Oliver Künzel, Jens Schmitt, Ralf Steinmetz
KOM Multimedia Communications
Darmstadt University of Technology
Merckstrasse 25, D-64283 Darmstadt, Germany
{zink, okuenzel, schmitt, steinmetz}@kom.tu-darmstadt.de

ABSTRACT

Layer encoded video is an elegant way to allow adaptive transmissions in the face of varying network conditions as well as it supports heterogeneity in networks and clients. As a drawback quality degradation can occur, caused by variations in the amount of transmitted layers. Recent work on reducing these variations makes assumptions about the perceived quality of those videos. The main goal of this paper respectively its motivation is to investigate the validity of these assumptions by subjective assessment. However, the paper is also an attempt to investigate fundamental issues for the human perception of layer encoded video with time-varying quality characteristics. For this purpose, we built a test environment for the subjective assessment of layer encoded video and conducted an empirical experiment in which 66 test candidates took part. The results of this subjective assessment are presented and discussed. To a large degree we were able to validate existing (unproven) assumptions about quality degradation caused by variations in layer encoded videos, however there were also some interesting, at first sight counterintuitive findings from our experiment.

Keywords

Empirical experiment, layer encoded video, human perception, video quality variations.

1. INTRODUCTION

1.1 Motivation

In the area of video streaming layer encoded video is an elegant way to overcome the inelastic characteristics of traditional video encoding formats like MPEG-1 or H.261. Layer encoded video is particularly useful in today's Internet where a lack of Quality of Service (QoS) mechanisms might make an adaptation to existing network conditions necessary. In addition, it bears the capability to support a large variety of clients while only a single file[1] has to be stored at a video server for each video object. The drawback of adaptive transmissions is the introduction of variations in the amount of transmitted layers during a streaming session. These variations affect the end-user's perceived quality and thus the acceptance of a service that is based on such technology.

Recent work that has focused on reducing those layer variations, either by employing intelligent buffering techniques at the client [2, 3, 4] or proxy caches [5, 6, 7] in the distribution network, made various assumptions about the perceived quality of videos with time-varying number of layers. To the best of our knowledge, these assumptions have not been verified by subjective assessment so far.

The lack of in-depth analysis about quality metrics for variations in layer encoded videos led us to conduct an empirical experiment based on subjective assessment to obtain results that can be used in classifying the perceived quality of such videos.

1.2 What is the Relation between Objective and Subjective Quality?

The goal of this research work is to investigate if general assumptions made about the quality metrics of variations in layer encoded videos can be verified by subjective assessment. We use the following example to explain our intention in more detail: A layer encoded video that is transmitted adaptively[2] to the client might have layer variations as shown in Figure 1. In Section 2.1 several quality metrics that allow the determination of the video's quality are presented. At first, we discuss the basics of these quality metrics. The most straightforward quality metric would be the total sum of all received segments (see Figure 1). However, common assumptions on the quality of a layer encoded video are that the quality is not only influenced by the total sum of received segments but also by the frequency of layer variations and the amplitude of those variations [3, 5, 7]. As shown in Figure 1 the amplitude specifies the height of a layer variation while the frequency determines the amount of layer variations.

All quality metrics we are aware of are based on these assumptions. Verifying all possible scenarios that are covered by those assumptions with an experiment based on subjective assessment is hard to achieve. Therefore, we decided to focus on basic scenarios that have the potential to answer the most fundamental questions, e.g., are the

[1] In contrast to the dynamic stream switching [1] approach where for each quality level one specific video file is required.

[2] Adaptively in this case means that the amount of layers transmitted to the client is based on some feedback from the network or the client, e.g., congestion control information.

sequences on the left in Figure 2 ((a1) and (b1)) more annoying than sequences on the right ((a2) and (b2)) for an end-user who views a corresponding video sequence. In this example, the first scenario ((a1) and (a2)) is focussed on the influence of the amplitude and the second ((b1) and (b2)) on the frequency of layer variations.

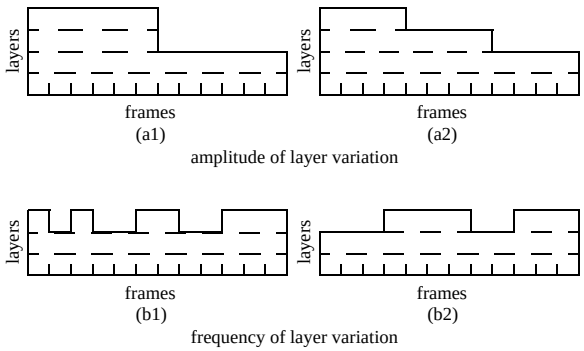


Figure 2. Quality criteria [3]

1.3 Outline

The paper is structured as follows. Section 2 reviews previous work on retransmission scheduling for layer encoded video and subjective assessment of video quality. The test environment and the subjective test method used for the experiment are described and discussed in Section 3. The details of the experimental setup are given in Section 4 and in Section 5 the results of the experiment are presented and discussed. Section 6 summarizes the major conclusions that can be drawn from the experiment.

2. RELATED WORK

The related work section is split in two parts since our work is influenced by the two research areas briefly surveyed in the following.

2.1 Retransmission Scheduling

The work presented in this paper has been motivated by our own work on quality improvement for layer encoded videos. During our investigation of favorable retransmission scheduling algorithms which are supposed to improve the quality of layer encoded videos stored on a cache [7], we realized that in related work the quality metrics for layer encoded videos are based on somewhat speculative assumptions only. To the best of our knowledge none of these assumptions is based on a subjective assessment.

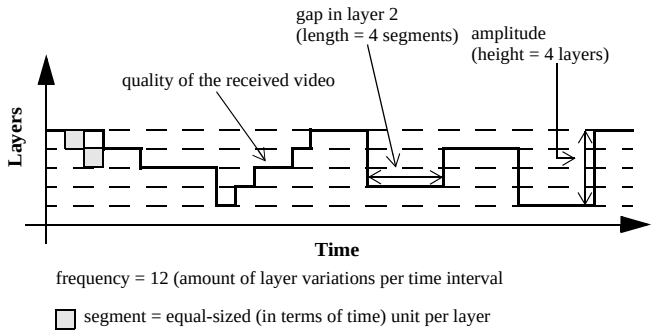


Figure 1. Quality of a layer encoded video at the client

In [3], Nelakuditi et al. state that a good metric should capture the amount of detail per frame as well as its uniformity across frames. I.e., if we compare the sequences of layers in a video shown in Figure 2 the quality of (a2) would be better than that of (a1) which is also valid for (b2) and (b1), according to their assumption. Their quality metric is based on the principle of giving a higher weight to lower layers and to longer runs of continuous frames in a layer.

The metric presented by the work of Rejaie et al. [5] is almost identical to the one advocated for in [3]. *Completeness* and *continuity* are the 2 parameters that are incorporated in this quality metric. *Completeness* of a layer is defined as the ratio of the layer size transmitted to its original (complete) size. E.g. the ratio of layer 2 in sequence (a1) in *Figure 2* would be 1 while the ratio for layer 3 would be 0.5. *Continuity* is the metric that covers the 'gaps' in a layer. It is defined as the average number of segments between two consecutive layer breaks (i.e., gaps). In contrast to the other metrics presented here, this metric is a per-layer metric.

In our previous work [7] we also made assumptions about the quality metrics for layer encoded videos. Similar to [3] we postulated that this metric should be based on a) the frequency of variations and b) the amplitude of variations.

2.2 Video Quality

There has been a substantial amount of research on methodologies for subjective assessment of video quality, e.g., [8] and [9], which contributed to form an ITU Recommendation on this issue [10]. This standard has been used as a basis for subjective assessment of encoders for digital video formats, in particular for MPEG-2 [11, 12] and MPEG-4 [13] but also on other standards like H.263+ [14]. The focus of interest for all these subjective assessment experiments was the quality of different coding and compression mechanisms. Our work, in contrast, is concerned with the quality degradation caused by variations in layer encoded videos. Like us, [15] is also concerned with layer encoded video and presents the results of an empirical evaluation of 4 hierarchical video encoding schemes. This is orthogonal to our work since the focus of their investigation is on the comparison between the different layered coding schemes and not on the human perception of layer variations.

In [16], a subjective quality assessment has been carried out in which the influence of the frame rate on the perceived quality is investigated. In contrast to our work elasticity in

the stream was achieved by frame rate variation and not by applying a layer encoded video format.

Effects of bit errors on the quality of MPEG-4 video were explored in [17] by subjective viewing measurements, but effects caused by layer variations were not examined.

Chen presents an investigation on an IP-based video conference system [18]. The focus in this work is mainly auditorium parameters like display size and viewing angle. A layer encoded video format is not used in this investigation.

Probably closest to our work, Lavington et al. [19] used an H.263+ two layer video format in their trial. In comparison to our approach, they were rather interested in the quality assessment of longer sequences (e.g., 25 min.). Instead of using identical pregenerated sequences that were presented to the test candidates, videos were streamed via an IP network to the clients and the quality was influenced in a fairly uncontrolled way by competing data originating from a traffic generator. The very specific goal of this work was to examine if reserving some of the network's bandwidth for either the base or the enhancement layer improves the perceived quality of the video, while we are rather interested on the influence of variations in layer encoded videos and try to verify some of the basic assumption made about the perceived quality in a subjective assessment experiment. Furthermore, we try to conduct this experiment in a much more controlled environment in order to achieve more significant and easier to interpret results.

3. TEST ENVIRONMENT

In this section, we first present the layer encoded video format used for the experiment, describe how we generated the test sequences, explain why we decided to use stimulus-comparison as the assessment method, and shortly present our test application.

3.1 Layer Encoded Video Format - SPEG

SPEG (Scalable MPEG) [20] is a simple modification to MPEG-1 which introduces scalability. In addition to the possibility of dropping complete frames (temporal scalability), which is already supported by MPEG-1 video, SNR scalability is introduced through layered quantization of DCT data [20]. The extension to MPEG-1 was made for two reasons. First, there are no freely available implementations of layered extensions for existing video standards (MPEG-2, MPEG-4), second, the granularity of scalability is improved by SPEG combining temporal and SNR scalability. As shown in Figure 3 a priority (p_0 - p_{11}) can be mapped to each layer. The QoS Mapper (see Figure 4, which depicts the SPEG pipeline and its components) uses the priority information to determine which layers are dropped and which are forwarded to the Net Streamer.

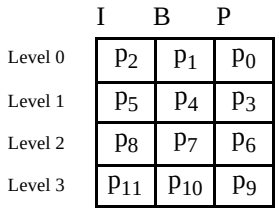


Figure 3. SPEG layer model

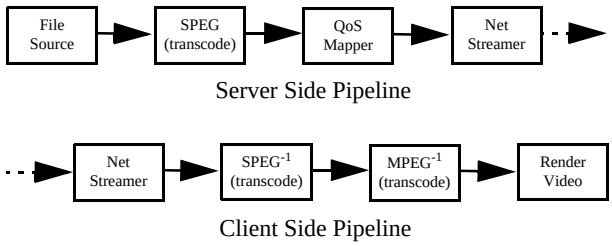


Figure 4. Pipeline for SPEG [21]

Our decision to use SPEG as a layer encoded video format is based on several reasons. SPEG is designed for a QoS-adaptive video-on-demand (VoD) approach, i.e., the data rate streamed to the client should be controlled by feedback from the network (e.g., congestion control information). In addition, the developers of SPEG also implemented a join function that re-transcodes SPEG into MPEG-1 [21] and therefore allows the use of standard MPEG-1 players, e.g., the Windows Media Player. We were not able to use scalable video encoders available as products (e.g., [22, 23]) because videos created by those can only be streamed to the corresponding clients which do neither allow the storage of the received data on a disk nor the creation of scheduled quality variations.

3.2 Test Generation - Full Control

Since our test sequences must be created in a deterministic manner, we slightly modified the SPEG pipeline. The most important difference is, that in our case data belonging to a certain layer must be dropped intentionally and not by an unpredictable feedback from the network or the client. This modification was necessary, since identical sequences must be presented to the test candidates in the kind of subjective assessment method that is used in our experiment. Therefore, we modified the QoS Mapper in a way that layers are dropped at certain points in time specified by manually created input data. We also added a second output path to the MPEG^{-1} module that allows us to write the resulting MPEG-1 data in a file.

3.3 Measurement Method -

Stimulus Comparison

The subjective assessment method is widely accepted for determining the perceived quality of images and videos. Research that was performed under the ITU-R lead to the development of a standard for such test methods [10]. The standard defines basically five different test methods double-stimulus impairment scale (DSIS), double-stimulus continuous quality-scale (DSCQS), single stimulus quality evaluation (SSCQE), simultaneous double stimulus for continuous evaluation (SDSCE), and stimulus-comparison (SC), respectively.

Since it was our goal to investigate the basic assumptions about the quality of layer encoded video, SSCQE and SDSCE are not the appropriate assessment method because comparisons between two videos are only possible on an identical time segment and not between certain intervals of the same video. In addition, SSCQE and SDSCE were designed to assess the quality of an encoder (e.g., MPEG-1) itself.

Two test methods which better suit the kind of investigations we want to perform are DSCQS and DSIS. Compared to SSCQE and SDSCE they allow to asses the quality of a codec in relation to data losses [8] and therefore, are more suitable if the impairment caused by the transmission path is investigated.

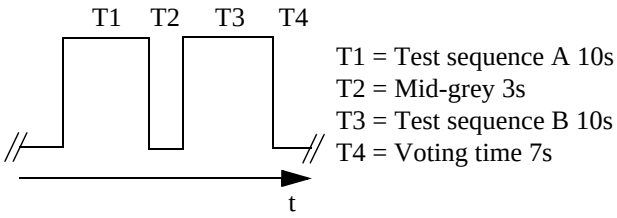


Figure 5. Presentation structure of test material

The SC method differs from DSCQS and DSIS in a way that two test sequences with unequal qualities are shown (see Figure 5) and the test candidates can vote on a scale as shown in Table 1. Comparing two impaired videos directly with each other is our primary goal. Since this is represented best by the SC method we decided to use this method in our test.

Additionally, preliminary tests have shown us that test candidates with experience in watching videos on a computer are less sensitive to impairment. I.e., they recognize the impairment but do not judge it as annoying as candidates who are unexperienced. This effect is dampened since only impaired sequences have to be compared with each other in a single test that is based on the SC method. Our preliminary tests with the DSIS method, where always the original sequence and an impaired sequence are compared, delivered results with less significance compared to tests performed with the SC method

Table 1: Comparison scale

Value	Compare
-3	much worse
-2	worse
-1	slightly worse
0	the same
1	slightly better
2	better
3	much better

3.4 Test Application - Enforcing Time Constraints

We created a small application[3] (see Figure 6) that allows an automated execution of the tests. Since we had to use a computer to present the videos anyway, we decided to let the candidates perform their voting also on the computer. Using this application has the advantage that we can easily enforce the time constraints demanded by the measurement method, because we allow voting only during a certain time interval. As a convenient side effect, the voting data is available in a machine readable format.

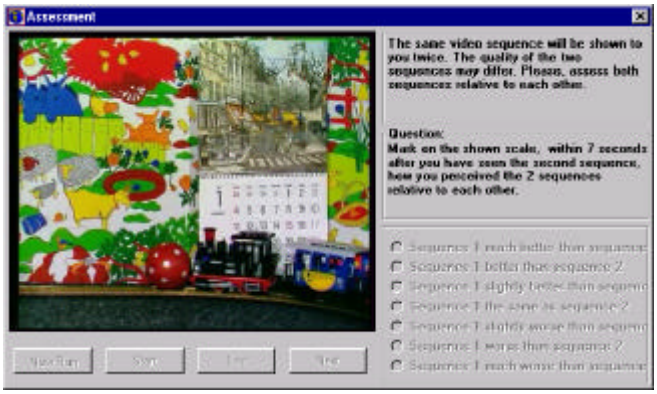


Figure 6. Application for experiment

4. EXPERIMENT

4.1 Scenario

Since quality metrics for layer encoded video are very general, we have to focus on some basic test cases in order to keep the amount of tests that should be performed in the experiment feasible. We decided to investigate isolated effects, one-by-one at a time, which on one hand keeps the

[3] A downloadable version of the test can be found at http://www.kom.e-technik.tu-darmstadt.de/video-assessment/

size of a test session reasonable and on the other hand still allows to draw conclusions for the general assumptions, as discussed above. That means we are rather interested in observing the quality ranking for isolated effects like frequency variations (as shown in sequences (b1) and (b2) in Figure 2) than for combined effects (as shown in Figure 1). This bears also the advantage that standardized test methods [10], which limit the sequence length to several seconds, can be applied. All patterns that were used for the experiment are shown in Figure 8.

4.2 Candidates

The experiment was performed with 66 test candidates (45 males and 21 females), between the age of 14 and 64. 55 of them had experiences with watching videos on a computer.

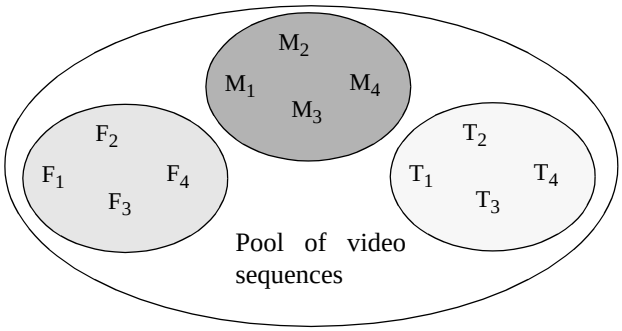
4.3 Procedure

Each candidate had to perform 15 different assessments, of which each single test lasted for 33 seconds. All 15 tests were executed according to the SC assessment method. The complete test session per candidate lasted for about 15 minutes[4], on average. We have chosen three video sequences for this experiment, that have been frequently used for subjective assessment [24]. The order of the 15 video sequences was changed randomly from candidate to candidate as proposed in the ITU-R B.500-10 standard [10] (see also Figure 7). After some initial questions (age, gender, profession) 3 assessments were executed as a warm-up phase. This should avoid that the test candidates are distracted by the content of the video sequences as reported by Aldridge et al. [11]. In order to avoid that two consecutive video sequences (e.g., F_2 is following F_1 immediately) have the same content we defined a pattern for the chronological order of the test sessions, as shown in Figure 7. F_x can be any video sequence from the F pool of sequences that has not been used in this specific test session, so far. Thus, a complete test session for a candidate could have a chronological order as shown for Figure 7.

4.4 Layer Patterns

Figure 8 shows the layer patterns of each single sequence that was used in the experiment, except for the first 3 warm-up tests where the comparison is performed between the first sequence that consists of 4 layers and the second that consists of only one layer. Each of the 3 groups shows the patterns that were used with one type of content. Comparisons were always performed between patterns that are shown in a row (e.g., (a1) and (a2)). As already mentioned in Section 1.2 it was our goal to examine fundamental assumptions about the influence of layer

[4] Only watching the sequences and voting took less time, but the candidates had as much time as they wanted to read the questions and possible answers for each test ahead of each test.



Pattern	I_1	I_2	I_3	F_x	M_x	T_x	F_x	M_x	T_x	F_x	M_x	T_x	F_x	M_x	T_x
Sequence 1	I_1	I_2	I_3	F_3	M_1	T_4	F_1	M_2	T_1	F_4	M_3	T_2	F_2	M_4	T_3
Sequence 2	I_1	I_2	I_3	F_3	M_1	T_4	F_1	M_2	T_1	F_4	M_3	T_2	F_2	M_4	T_3

F = Farm
M = Mobile & Calendar
T = Table Tennis

Figure 7. Random generation of test sequence order

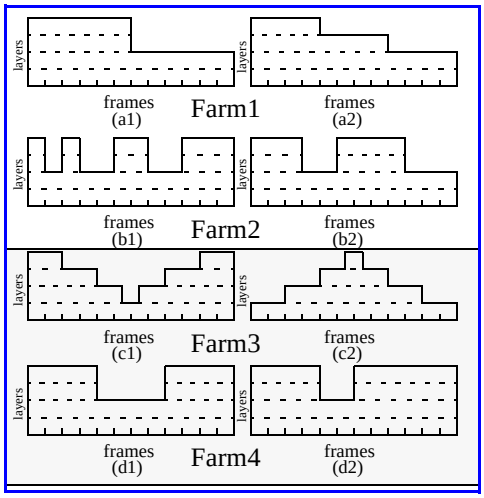
changes on perceived quality. This is also reflected by the kind of patterns we decided to use in the experiment. It must be mentioned that the single layers are not equal in size (contrary to the presentation in Section 8). The size of the n^{th} layer is rather determined by the following expression: $s_n = 2s_{n-1}$. Thus, segments of different layers have different sizes. Preliminary experiments have shown that equal layer sizes are not appropriate to make layer changes perceivable. Since there exist layered schemes that produce layers with sizes similar to ours [25, 26], we regard this a realistic assumption.

In the experiment, we differentiate between two groups of tests, i.e., one group in which the amount of segments used by a pair of sequences is equal and one in which the amount differs (the latter has a shaded background in Figure 8). We made this distinction because we are mainly interested in how the result of this experiment could be used to improve the retransmission scheduling technique (see Section 2.1) where it is necessary to compare the influence of additional segments that is added on different locations in a sequence.

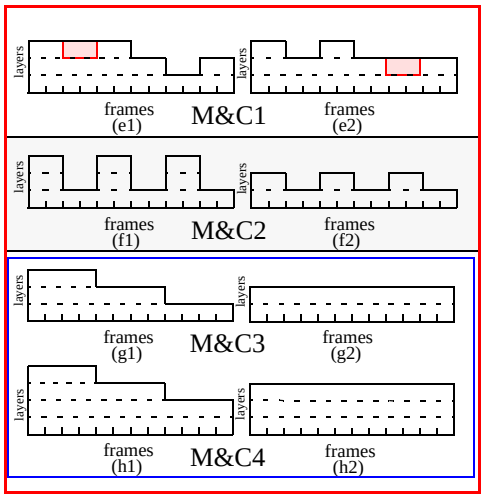
Since segments from different layers are not equal in size, the amount of data for the compared sequences differs. However, somewhat surprisingly, as we discuss in Section 5.3, a larger amount of data does not necessarily lead to a better perceived quality. Additional tests with different quantities of segments in between a pair were chosen to answer additional questions and make the experiment more consistent as we show in Section 5.2.

5. RESULTS

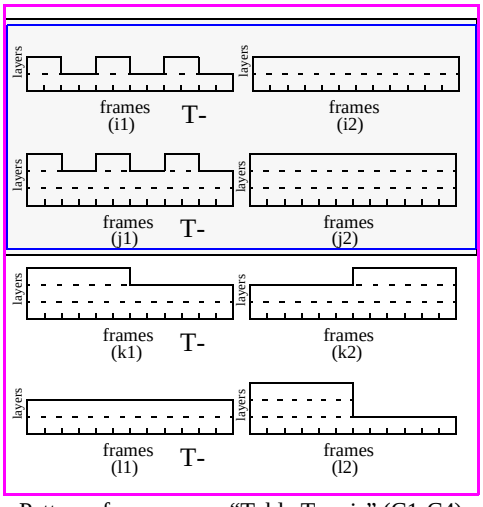
In this section, we present the results of the experiment described in Section 4. Since we analyze the gathered data statistically it must clearly be mentioned that the presented results cannot prove an assumption but only make it less or



Patterns for Sequence “Farm” (F1-F4)



Patterns for sequence “Mobile & Calendar” (M1-M4)



Patterns for sequence “Table Tennis” (C1-C4)

Figure 8. Segments that were compared in the experiment

more likely based on the gathered data. The overall results of all experiments are summarized in Figure 9 and are discussed in the following subsections.

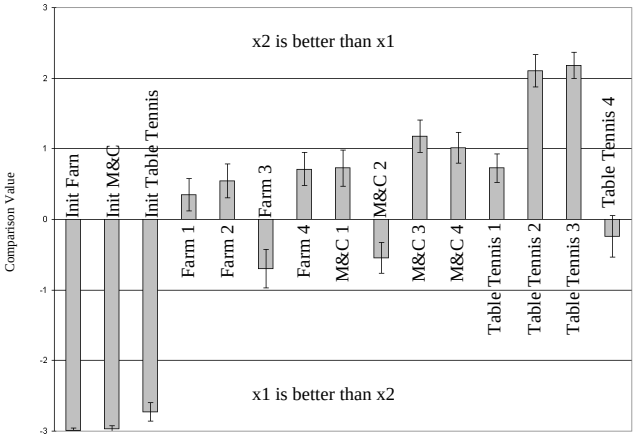


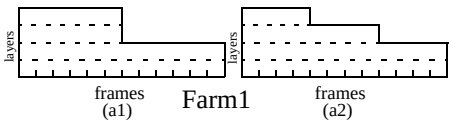
Figure 9. Average and 95% confidence interval for the different tests of the experiment

5.1 Same Amount of Segments

In this section, we discuss the results for the assessments of tests in which the total sum of segments is equal. That means the space covered by the pattern of both sequences is identical.

5.1.1 Farm1: Amplitude

In this assessment the stepwise decrease was rated slightly better than one single but higher decrease. The result shows a tendency that the assumptions that were made about the amplitude of a layer change (as described in Section 2.1) are correct.



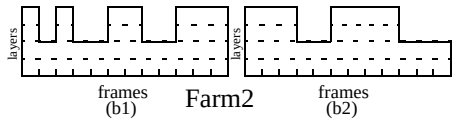
Average:0.35 Conf.-Interval Size:0.23

Figure 10. Farm1

5.1.2 Farm2: Frequency

The result of this test has an even higher likelihood that the second sequence has a better perceived quality than it is the case for *Farm1*. It tends to confirm the assumption that the frequency of layer changes influences the perceived quality,

since, on average, test candidates ranked the quality of the sequence with lesser layer changes better.

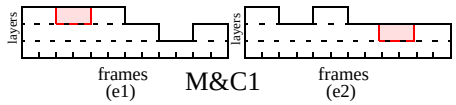


Average:0.55 Conf.-Interval Size:0.24

Figure 11. Farm2

5.1.3 M&C1: Closing the gap

This test should try to answer the question, if it would be better to close a gap in a layer on a higher or lower level. The majority of the test candidates decided that filling the gap on a lower level results in a better quality than otherwise. This result tends to affirm our assumptions made for retransmission scheduling in [7].

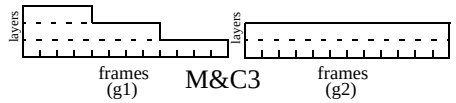


Average:0.73 Conf.-Interval Size:0.25

Figure 12. M&C1

5.1.4 M&C3: Constancy

Even more significant than in the preceding tests, the candidates favored the sequence with no layer changes as the one with the better quality. One may judge this a trivial and unnecessary test, but from our point of view the result is not that obvious, since (g1) starts with a higher amount of layers. The outcome of this test implies that it might be better, in terms of perceived quality, to transmit less but a constant amount of layers.

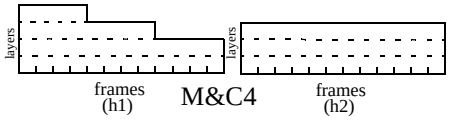


Average:1.18 Conf.-Interval Size:0.23

Figure 13. M&C3

5.1.5 M&C4: Constancy at a higher level

This test was to examine if an increase of the overall level (in this case by comparison to Section 5.1.4) has an influence on the perceived quality. Comparing the results of both tests (*M&C3* and *M&C4*) shows no significant change in the test candidates' assessment. 66% of the test candidates judge the second sequences ((g2) and (h2)) better (values 1-3 in Table 1) in both cases which makes it likely that the overall level has no influence on the perceived quality.

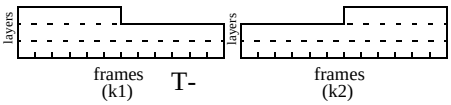


Average:1.02 Conf.-Interval Size:0.22

Figure 14. M&C4

5.1.6 Tennis3: All is well that ends well

The result of this test shows the tendency that increasing the amount of layers in the end leads to a higher perceived quality. The result is remarkably strong (the highest bias of all tests). Future tests, that will be of longer duration and executed in a different order (first (k2) than (k1)), will show how the memory-effect [11] of the candidates influenced this test.

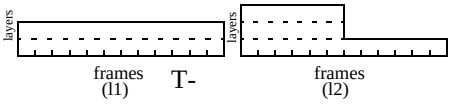


Average:2.18 Conf.-Interval Size:0.19

Figure 15. Tennis3

5.1.7 Tennis4: The exception proves the rule

This test is the only one out of the 12 tests in which the 95% confidence interval covers both areas (better, worth) of the judgement scale. If we regard the average only, the result is a little bit surprising since it contradicts the results from Section 5.1.1 and Section 5.1.4, respectively. At this stage of the investigation, we can only assume that also the content might have an influence on the perceived quality. But to gain more insight in this phenomenon further experiments are necessary.



Average:-0.24 Conf.-Interval Size:0.29

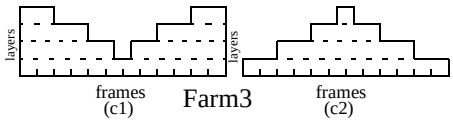
Figure 16. Tennis4

5.2 Different amount of Segments

In the following 5 tests the total amount of segments per sequence differs. All 5 tests have in common that the perceived quality of the sequence consisting of a pattern that covers a larger number of segments were ranked better.
This is obvious, but it makes the overall result more consistent, because test candidates mostly realized this quality difference.

5.2.1 Farm3: Decrease vs. increase

Starting with a higher amount of layers, decreasing the amount of layers, and increasing the amount of layers in the end again seems to provide a better perceivable quality than starting with a low amount of layers, increasing this amount of layers, and going back to a low amount of layers at the end of the sequence. This might be caused by the fact that test candidates are very concentrated in the beginning and the end of the sequence and that, in the first case details become clear right in the beginning of the sequence.

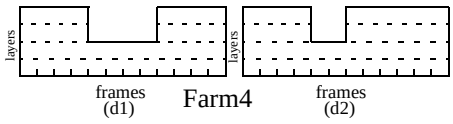


Average:-0.70 Conf.-Interval Size:0.27

Figure 17. Farm3

5.2.2 Farm4: Keep the gap small

In this test, it was our goal to investigate how the size of a gap may influence the perceived quality. The majority of test candidates (37 out of 66) judged the quality of the sequence with a smaller gap slightly better (Only 5 out of 66 judged the first sequence better). This indicates that filling a gap partly can be beneficial.

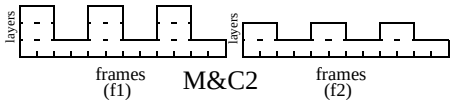


Average:0.71 Conf.-Interval Size:0.23

Figure 18. Farm4

5.2.3 M&C2: Increasing the amplitude

The effect of the amplitude height should be investigated in this test. The result shows that, in contrast to existing assumptions (see Section 2.1), an increased amplitude can lead to a better perceived quality.

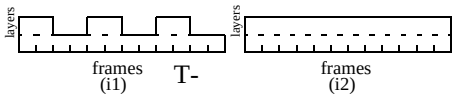


Average:-0.55 Conf.-Interval Size:0.22

Figure 19. M&C2

5.2.4 Tennis1: Closing all gaps

This test is contrary to *M&C2* where the additional segments are used to close the existing gaps instead of increasing the amplitude of already better parts of the sequence. This strategy decreases the frequency of layer changes. Test candidates, on average, judged the sequence without layer changes better. The result of this test reaffirms the tendency that was already noticed in Section 5.1.2, that the perceived quality is influenced by the frequency of layer changes. If we carefully compare the results of *M&C2* and *Tennis1*, a tendency towards filling the gaps and thus decreasing the frequency instead of increasing the amount of already increased parts of the sequence is recognizable. Definitely, further investigations are necessary to confirm this tendency, because, here, the results of tests with different contents are compared and we have not investigated the influence of the content on the perceived quality, so far.

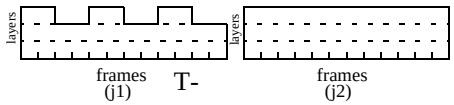


Average:0.73 Conf.-Interval Size:0.20

Figure 20. Tennis1

5.2.5 Tennis2: Closing all gaps at a higher level

In comparison to *Tennis1*, here, we were interested in how an overall increase of the layers (in this case by one layer) would influence the test candidates judgement. Again the sequence with no layer changes is judged better, even with a higher significance than for *Tennis1*. This might be caused by the fact that the amount of layer is higher in general in *Tennis2*.



Average:2.11 Conf.-Interval Size:0.23

Figure 21. Tennis1

5.3 Sequence Size and Quality

As already mentioned in Section 4.4, segments of different layers are not equal in size. Hence, the data size for patterns with an equal amount of segments might not be identical. Here we give an example that shows that sequences consisting of a higher amount of data are not necessarily judged better concerning the perceived quality of the sequence. If we compare, e.g., the tests *M&C1, M&C3,* and *M&C4,* respectively, we recognize that the storage size of the first sequence is always larger than the one of the second. This relation is shown in Figure 22. The results of our experiment show that the average judgement is in contrast to the storage size (Figure 8) where the second sequence has, according to the test candidates, a better perceived quality.

6. CONCLUSION

In this paper, we presented the results of an empirical experiment based on subjective assessment of variations in layer encoded video. A statistical analysis of the experiment

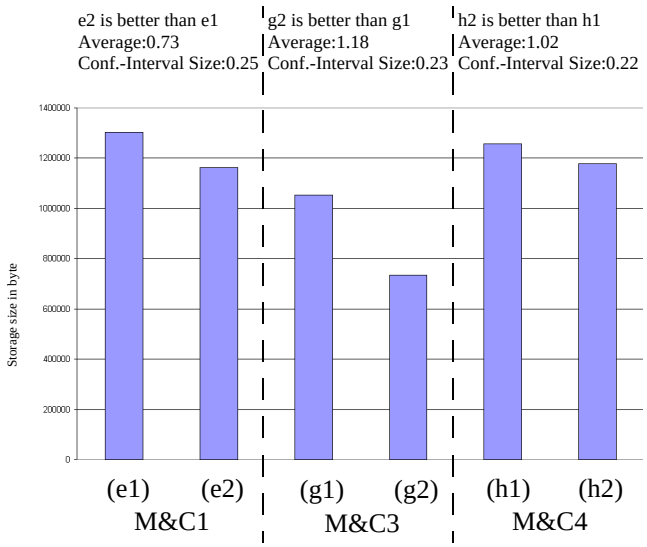


Figure 22. Storage size of sequences

mostly validates assumptions that were made in relation to layer variations and the perceived quality of a video:

- The frequency of variations should be kept as small as possible.
- If a variation can not be avoided the amplitude of the variation should be kept as small as possible.

One basic conclusion from the results in Section 5.2 is: adding information to a layered video increases its average quality. But, as we already assumed in our work on retransmission scheduling, adding information at different locations can have a substantial effect on the perceived quality. Assumptions we made for our heuristics in retransmission scheduling (as well as others' assumptions) could be substantiated by this investigation (see Section 5.1). That means, it is more likely that the perceived quality of a layer encoded video is improved if

- the lowest quality level is increased, and
- gaps in lower layers are filled.

The results from Section 5.3 should be used to refine the retransmission scheduling heuristics in relation to the size of each single layer. Therefore, the metric that represents the quality improvement must also take into account that it might be more expensive to retransmit a segment of layer n+1 than of layer n. Another interesting outcome of the experiment is the fact that a quality improvement may be achieved by retransmitting less data (Section 5.1.3), if a layered encoding scheme is used in which the layers are not of identical size. The obtained results can, in addition, be used to refine caching replacement policies that operate on a layer level [5] as well as layered multicast transmission schemes which try to offer heterogeneous services to different subscribers as, e.g., in the receiver-driven layered multicast RLM [27] scheme and its derivations.

The results of this investigation clearly strengthen the assumption that a differentiation between objective and subjective quality, in the case of variations in layer encoded video, must be made.

Nevertheless, it must be admitted that the presented work is only an initial investigation in the subjective impression of variations in layer encoded videos. In further work, we want to explore sequences with a longer duration (up to several minutes). In a next step, we will investigate if the shown sequences can be combined and if the subjective assessment is still consistent with the separated results. E.g., in this experiment sequences (e2) and (g2) were judged better than (e1) and (g1), will a sequence that combines (e2) and (g2) also be judged better than a sequence that combines (e1) and (g1)? We are also interested in how the content of a sequence influences the perceived quality.

7. ACKNOWLEDGMENTS

The authors would like to thank Rico Tunk for creating the test application, Charles "Buck" Krasic for his support on SPEG, the test candidates for taking the time to perform the assessment, and RTL Television for providing the video sequences.

8. REFERENCES

[1] J. Lu. Signal Processing for Internet Video Streaming: A Review. In *Proceedings of SPIE Image and Video Communications and Processing, San Jose, CA, USA*. SPIE - The International Society for Optical Engineering, January 2000.

[2] D. Saparilla and K. W. Ross. Optimal Streaming of Layered Video. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies 2000 (INFOCOM'00), Tel-Aviv, Israel*, pages 737–746, March 2000.

[3] S. Nelakuditi, R. R. Harinath, E. Kusmierek, and Z.-L. Zhang. Providing Smoother Quality Layered Video Stream. In *Proceedings of the 10th International Workshop on Network and Operating System Support for Digital Audio and Video, Raleigh, NC, USA*, June 2000.

[4] R. Rejaie, M. Handley, and D. Estrin. Quality Adaptation for Congestion Controlled Video Playback over the Internet. In *Proceedings of the ACM SIGCOMM '99 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication 1999, New York, NY, USA*, pages 189–200, August 1999.

[5] R. Rejaie, H. Yu, M. Handley, and D. Estrin. Multimedia Proxy Caching for Quality Adaptive Streaming Applications in the Internet. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies 2000 (INFOCOM'00), Tel-Aviv, Israel*, pages 980–989, March 2000.

[6] R. Rejaie and J. Kangasharju. Mocha: A Quality Adaptive Multimedia Proxy Cache for Internet Streaming. In *Proceedings of the 11th International Workshop on Network and Operating System Support for Digital Audio and Video, Port Jefferson, New York, USA*, pages 3–10, June 2001.

[7] M. Zink, J. Schmitt, and R. Steinmetz. Retransmission Scheduling in Layered Video Caches, April 2002. Accepted at ICC 2002, New York, New York, USA.

[8] T. Alpert and J.-P. Evain. Subjective quality evaluation - The SSCQE and DSCQE methodologies. EBU Technical Review, February 1997.

[9] E. Kohler, M. Handley, S. Floyd, and J. Padhye. Datagram Control Protocol (DCP). Internet Draft, November 2001. Work in Progress.

[10] ITU-R: Methodology for the Subjective Assessment of the Quality of Television Picture. International Standard, 2000. ITU-R BT.500-10.

[11] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson. Measurement of scene-dependent quality variations in digitally coded television pictures. *IEE Proceedings on Vision, Image and Signal Processing*, 142, 3:149–154, 1995.

[12] R. Aldridge, D. Hands, D. Pearson, and N. Lodge. Continuous quality assessment of digitally-coded television pictures. *IEE Proceedings on Vision, Image and Signal Processing*, 145, 2:116–123, 1998.

[13] F. Pereira and T. Alpert. MPEG-4 video subjective test procedures and results. *IEEE Transactions on Circuits and Systems for Video Technology*, 7, 1:32–51, 1997.

[14] M. Masry and S. Hemami. An analysis of subjective quality in low bit rate video. In *International Conference on Image Processing (ICIP), 2001, Thessaloniki, Greece*, pages 465–468. IEEE Computer Society Press, October 2001.

[15] C. Kuhmünch and C. Schremmer. Empirical Evaluation of Layered Video Coding Schemes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece*, pages 1013–1016, October 2001.

[16] T. Hayashi, S.Yamasaki, N. Morita, H. Aida, M. Takeichi, and N. Doi. Effects of IP packet loss and picture frame reduction on MPEG1 subjective quality. In *3rd Workshop on Multimedia Signal Processing, Copenhagen, Denmark*, pages 515–520. IEEE Computer Society Press, September 1999.

[17] S. Gringeri, R. Egorov, K. Shuaib, A. Lewis, and B. Basch. Robust compression and transmission of MPEG-4 video. In *Proceedings of the ACM Multimedia Conference 1999, Orlando, Florida, USA*, pages 113–120, October 1999.

[18] M. Chen. Design of a virtual auditorium. In *Proceedings of the ACM Multimedia Conference 2001, Ottawa, Canada*, pages 19–28, September 2001.

[19] S. Lavington, N. Dewhurst, and M. Ghanbari. The Performance of Layered Video over an IP Network. *Signal Processing: Image Communication, Elsevier Science*, 16, 8:785–794, 2001.

[20] C. Krasic and J. Walpole. Priority-Progress Streaming for Quality-Adaptive Multimedia. In *ACM Multimedia Doctoral Symposium, Ottawa, Canada*, October 2001.

[21] C. Krasic and J. Walpole. QoS Scalability for Streamed Media Delivery. Technical Report OGI CSE Technical Report CSE-99-011, Oregon Graduate Institute of Science & Technology, September 1999.

[22] Intel. Developers - What Intel Streaming Web Video Software Can Do For You, 2000. http://developer.intel.com/ial/swv/developer.htm.

[23] PacketVideo. Technical White Paper: PacketVideo Multimedia Technology Overview, 2001. http://www.packetvideo.com/pdf/pv_whitepaper.pdf.

[24] R. Neff and A. Zakhor. Matching Pursuit Video Coding–Part I: Dictionary Approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 12, 1:13–26, 2002.

[25] J. Hartung, A. Jacquin, J. Pawlyk, and K. Shipley. A Real-time Scalable Software Video Codec for Collaborative Applications over Packet Networks. In *Proceedings of the ACM Multimedia Conference 1998, Britol, UK*, pages 419–426, September 1998.

[26] L. Vicisano, L. Rizzo, and J. Crowcroft. TCP-like congestion control for layered multicast data transfer. In *Proceedings of the 17th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'98)*, pages 996–1003. IEEE Computer Society Press, March 1998.

[27] S. McCanne, M. Vetterli, and V. Jacobson. Receiver-driven layered multicast. In *Proceedings of ACM SIGCOMM'96*, Palo Alto, CA, August 1996.

Subjective Impression of Variations in Layer Encoded Videos

Michael Zink, Oliver Künzel, Jens Schmitt, Ralf Steinmetz
KOM Multimedia Communications
Darmstadt University of Technology
Merckstrasse 25, D-64283 Darmstadt, Germany
{zink, okuenzel, schmitt, steinmetz}@kom.tu-darmstadt.de

ABSTRACT

Layer encoded video is an elegant way to allow adaptive transmissions in the face of varying network conditions as well as it supports heterogeneity in networks and clients. As a drawback quality degradation can occur, caused by variations in the amount of transmitted layers. Recent work on reducing these variations makes assumptions about the perceived quality of those videos. The main goal of this paper respectively its motivation is to investigate the validity of these assumptions by subjective assessment. However, the paper is also an attempt to investigate fundamental issues for the human perception of layer encoded video with time-varying quality characteristics. For this purpose, we built a test environment for the subjective assessment of layer encoded video and conducted an empirical experiment in which 66 test candidates took part. The results of this subjective assessment are presented and discussed. To a large degree we were able to validate existing (unproven) assumptions about quality degradation caused by variations in layer encoded videos, however there were also some interesting, at first sight counterintuitive findings from our experiment.

Keywords

Empirical experiment, layer encoded video, human perception, video quality variations.

1. INTRODUCTION

1.1 Motivation

In the area of video streaming layer encoded video is an elegant way to overcome the inelastic characteristics of traditional video encoding formats like MPEG-1 or H.261. Layer encoded video is particularly useful in today's Internet where a lack of Quality of Service (QoS) mechanisms might make an adaptation to existing network conditions necessary. In addition, it bears the capability to support a large variety of clients while only a single file[1] has to be stored at a video server for each video object. The drawback of adaptive transmissions is the introduction of variations in the amount of transmitted layers during a streaming session. These variations affect the end-user's perceived quality and thus the acceptance of a service that is based on such technology.

Recent work that has focused on reducing those layer variations, either by employing intelligent buffering techniques at the client [2, 3, 4] or proxy caches [5, 6, 7] in the distribution network, made various assumptions about the perceived quality of videos with time-varying number of layers. To the best of our knowledge, these assumptions have not been verified by subjective assessment so far.

The lack of in-depth analysis about quality metrics for variations in layer encoded videos led us to conduct an empirical experiment based on subjective assessment to obtain results that can be used in classifying the perceived quality of such videos.

1.2 What is the Relation between Objective and Subjective Quality?

The goal of this research work is to investigate if general assumptions made about the quality metrics of variations in layer encoded videos can be verified by subjective assessment. We use the following example to explain our intention in more detail: A layer encoded video that is transmitted adaptively[2] to the client might have layer variations as shown in Figure 1. In Section 2.1 several quality metrics that allow the determination of the video's quality are presented. At first, we discuss the basics of these quality metrics. The most straightforward quality metric would be the total sum of all received segments (see Figure 1). However, common assumptions on the quality of a layer encoded video are that the quality is not only influenced by the total sum of received segments but also by the frequency of layer variations and the amplitude of those variations [3, 5, 7]. As shown in Figure 1 the amplitude specifies the height of a layer variation while the frequency determines the amount of layer variations.

All quality metrics we are aware of are based on these assumptions. Verifying all possible scenarios that are covered by those assumptions with an experiment based on subjective assessment is hard to achieve. Therefore, we decided to focus on basic scenarios that have the potential to answer the most fundamental questions, e.g., are the

[1] In contrast to the dynamic stream switching [1] approach where for each quality level one specific video file is required.

[2] Adaptively in this case means that the amount of layers transmitted to the client is based on some feedback from the network or the client, e.g., congestion control information.

sequences on the left in Figure 2 ((a1) and (b1)) more annoying than sequences on the right ((a2) and (b2)) for an end-user who views a corresponding video sequence. In this example, the first scenario ((a1) and (a2)) is focussed on the influence of the amplitude and the second ((b1) and (b2)) on the frequency of layer variations.

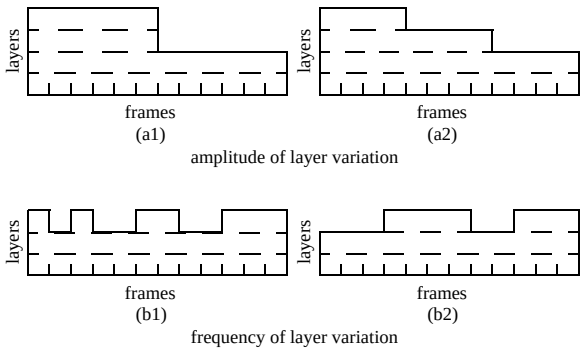


Figure 2. Quality criteria [3]

1.3 Outline

The paper is structured as follows. Section 2 reviews previous work on retransmission scheduling for layer encoded video and subjective assessment of video quality. The test environment and the subjective test method used for the experiment are described and discussed in Section 3. The details of the experimental setup are given in Section 4 and in Section 5 the results of the experiment are presented and discussed. Section 6 summarizes the major conclusions that can be drawn from the experiment.

2. RELATED WORK

The related work section is split in two parts since our work is influenced by the two research areas briefly surveyed in the following.

2.1 Retransmission Scheduling

The work presented in this paper has been motivated by our own work on quality improvement for layer encoded videos. During our investigation of favorable retransmission scheduling algorithms which are supposed to improve the quality of layer encoded videos stored on a cache [7], we realized that in related work the quality metrics for layer encoded videos are based on somewhat speculative assumptions only. To the best of our knowledge none of these assumptions is based on a subjective assessment.

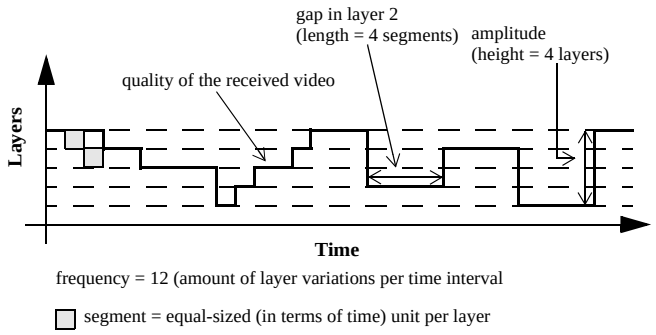


Figure 1. Quality of a layer encoded video at the client

In [3], Nelakuditi et al. state that a good metric should capture the amount of detail per frame as well as its uniformity across frames. I.e., if we compare the sequences of layers in a video shown in Figure 2 the quality of (a2) would be better than that of (a1) which is also valid for (b2) and (b1), according to their assumption. Their quality metric is based on the principle of giving a higher weight to lower layers and to longer runs of continuous frames in a layer.

The metric presented by the work of Rejaie et al. [5] is almost identical to the one advocated for in [3]. *Completeness* and *continuity* are the 2 parameters that are incorporated in this quality metric. *Completeness* of a layer is defined as the ratio of the layer size transmitted to its original (complete) size. E.g. the ratio of layer 2 in sequence (a1) in *Figure 2* would be 1 while the ratio for layer 3 would be 0.5. *Continuity* is the metric that covers the 'gaps' in a layer. It is defined as the average number of segments between two consecutive layer breaks (i.e., gaps). In contrast to the other metrics presented here, this metric is a per-layer metric.

In our previous work [7] we also made assumptions about the quality metrics for layer encoded videos. Similar to [3] we postulated that this metric should be based on a) the frequency of variations and b) the amplitude of variations.

2.2 Video Quality

There has been a substantial amount of research on methodologies for subjective assessment of video quality, e.g., [8] and [9], which contributed to form an ITU Recommendation on this issue [10]. This standard has been used as a basis for subjective assessment of encoders for digital video formats, in particular for MPEG-2 [11, 12] and MPEG-4 [13] but also on other standards like H.263+ [14]. The focus of interest for all these subjective assessment experiments was the quality of different coding and compression mechanisms. Our work, in contrast, is concerned with the quality degradation caused by variations in layer encoded videos. Like us, [15] is also concerned with layer encoded video and presents the results of an empirical evaluation of 4 hierarchical video encoding schemes. This is orthogonal to our work since the focus of their investigation is on the comparison between the different layered coding schemes and not on the human perception of layer variations.

In [16], a subjective quality assessment has been carried out in which the influence of the frame rate on the perceived quality is investigated. In contrast to our work elasticity in

the stream was achieved by frame rate variation and not by applying a layer encoded video format.

Effects of bit errors on the quality of MPEG-4 video were explored in [17] by subjective viewing measurements, but effects caused by layer variations were not examined.

Chen presents an investigation on an IP-based video conference system [18]. The focus in this work is mainly auditorium parameters like display size and viewing angle. A layer encoded video format is not used in this investigation.

Probably closest to our work, Lavington et al. [19] used an H.263+ two layer video format in their trial. In comparison to our approach, they were rather interested in the quality assessment of longer sequences (e.g., 25 min.). Instead of using identical pregenerated sequences that were presented to the test candidates, videos were streamed via an IP network to the clients and the quality was influenced in a fairly uncontrolled way by competing data originating from a traffic generator. The very specific goal of this work was to examine if reserving some of the network's bandwidth for either the base or the enhancement layer improves the perceived quality of the video, while we are rather interested on the influence of variations in layer encoded videos and try to verify some of the basic assumption made about the perceived quality in a subjective assessment experiment. Furthermore, we try to conduct this experiment in a much more controlled environment in order to achieve more significant and easier to interpret results.

3. TEST ENVIRONMENT

In this section, we first present the layer encoded video format used for the experiment, describe how we generated the test sequences, explain why we decided to use stimulus-comparison as the assessment method, and shortly present our test application.

3.1 Layer Encoded Video Format - SPEG

SPEG (Scalable MPEG) [20] is a simple modification to MPEG-1 which introduces scalability. In addition to the possibility of dropping complete frames (temporal scalability), which is already supported by MPEG-1 video, SNR scalability is introduced through layered quantization of DCT data [20]. The extension to MPEG-1 was made for two reasons. First, there are no freely available implementations of layered extensions for existing video standards (MPEG-2, MPEG-4), second, the granularity of scalability is improved by SPEG combining temporal and SNR scalability. As shown in Figure 3 a priority (p_0 - p_{11}) can be mapped to each layer. The QoS Mapper (see Figure 4, which depicts the SPEG pipeline and its components) uses the priority information to determine which layers are dropped and which are forwarded to the Net Streamer.

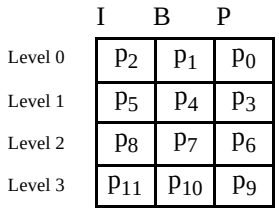


Figure 3. SPEG layer model

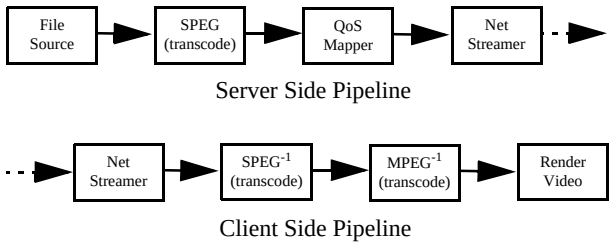


Figure 4. Pipeline for SPEG [21]

Our decision to use SPEG as a layer encoded video format is based on several reasons. SPEG is designed for a QoS-adaptive video-on-demand (VoD) approach, i.e., the data rate streamed to the client should be controlled by feedback from the network (e.g., congestion control information). In addition, the developers of SPEG also implemented a join function that re-transcodes SPEG into MPEG-1 [21] and therefore allows the use of standard MPEG-1 players, e.g., the Windows Media Player. We were not able to use scalable video encoders available as products (e.g., [22, 23]) because videos created by those can only be streamed to the corresponding clients which do neither allow the storage of the received data on a disk nor the creation of scheduled quality variations.

3.2 Test Generation - Full Control

Since our test sequences must be created in a deterministic manner, we slightly modified the SPEG pipeline. The most important difference is, that in our case data belonging to a certain layer must be dropped intentionally and not by an unpredictable feedback from the network or the client. This modification was necessary, since identical sequences must be presented to the test candidates in the kind of subjective assessment method that is used in our experiment. Therefore, we modified the QoS Mapper in a way that layers are dropped at certain points in time specified by manually created input data. We also added a second output path to the MPEG^{-1} module that allows us to write the resulting MPEG-1 data in a file.

3.3 Measurement Method -

Stimulus Comparison

The subjective assessment method is widely accepted for determining the perceived quality of images and videos. Research that was performed under the ITU-R lead to the development of a standard for such test methods [10]. The standard defines basically five different test methods double-stimulus impairment scale (DSIS), double-stimulus continuous quality-scale (DSCQS), single stimulus quality evaluation (SSCQE), simultaneous double stimulus for continuous evaluation (SDSCE), and stimulus-comparison (SC), respectively.

Since it was our goal to investigate the basic assumptions about the quality of layer encoded video, SSCQE and SDSCE are not the appropriate assessment method because comparisons between two videos are only possible on an identical time segment and not between certain intervals of the same video. In addition, SSCQE and SDSCE were designed to assess the quality of an encoder (e.g., MPEG-1) itself.

Two test methods which better suit the kind of investigations we want to perform are DSCQS and DSIS. Compared to SSCQE and SDSCE they allow to asses the quality of a codec in relation to data losses [8] and therefore, are more suitable if the impairment caused by the transmission path is investigated.

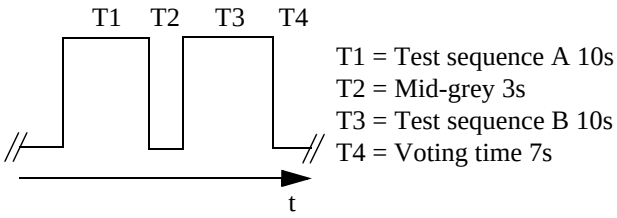


Figure 5. Presentation structure of test material

The SC method differs from DSCQS and DSIS in a way that two test sequences with unequal qualities are shown (see Figure 5) and the test candidates can vote on a scale as shown in Table 1. Comparing two impaired videos directly with each other is our primary goal. Since this is represented best by the SC method we decided to use this method in our test.

Additionally, preliminary tests have shown us that test candidates with experience in watching videos on a computer are less sensitive to impairment. I.e., they recognize the impairment but do not judge it as annoying as candidates who are unexperienced. This effect is dampened since only impaired sequences have to be compared with each other in a single test that is based on the SC method. Our preliminary tests with the DSIS method, where always the original sequence and an impaired sequence are compared, delivered results with less significance compared to tests performed with the SC method

Table 1: Comparison scale

Value	Compare
-3	much worse
-2	worse
-1	slightly worse
0	the same
1	slightly better
2	better
3	much better

3.4 Test Application - Enforcing Time Constraints

We created a small application[3] (see Figure 6) that allows an automated execution of the tests. Since we had to use a computer to present the videos anyway, we decided to let the candidates perform their voting also on the computer. Using this application has the advantage that we can easily enforce the time constraints demanded by the measurement method, because we allow voting only during a certain time interval. As a convenient side effect, the voting data is available in a machine readable format.

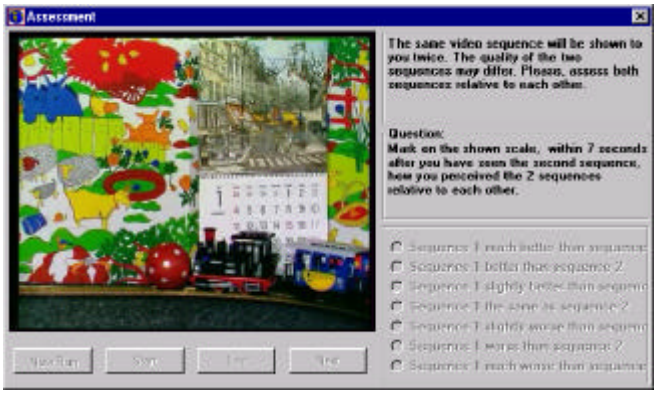


Figure 6. Application for experiment

4. EXPERIMENT

4.1 Scenario

Since quality metrics for layer encoded video are very general, we have to focus on some basic test cases in order to keep the amount of tests that should be performed in the experiment feasible. We decided to investigate isolated effects, one-by-one at a time, which on one hand keeps the

[3] A downloadable version of the test can be found at http://www.kom.e-technik.tu-darmstadt.de/video-assessment/

size of a test session reasonable and on the other hand still allows to draw conclusions for the general assumptions, as discussed above. That means we are rather interested in observing the quality ranking for isolated effects like frequency variations (as shown in sequences (b1) and (b2) in Figure 2) than for combined effects (as shown in Figure 1). This bears also the advantage that standardized test methods [10], which limit the sequence length to several seconds, can be applied. All patterns that were used for the experiment are shown in Figure 8.

4.2 Candidates

The experiment was performed with 66 test candidates (45 males and 21 females), between the age of 14 and 64. 55 of them had experiences with watching videos on a computer.

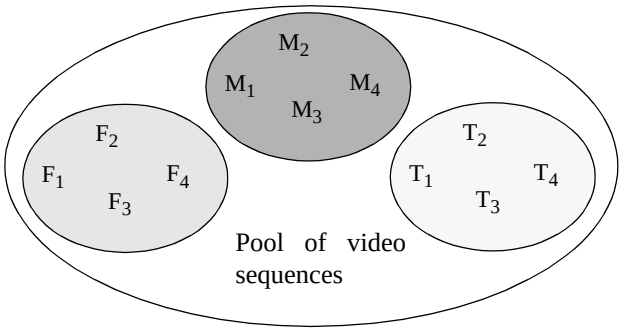
4.3 Procedure

Each candidate had to perform 15 different assessments, of which each single test lasted for 33 seconds. All 15 tests were executed according to the SC assessment method. The complete test session per candidate lasted for about 15 minutes[4], on average. We have chosen three video sequences for this experiment, that have been frequently used for subjective assessment [24]. The order of the 15 video sequences was changed randomly from candidate to candidate as proposed in the ITU-R B.500-10 standard [10] (see also Figure 7). After some initial questions (age, gender, profession) 3 assessments were executed as a warm-up phase. This should avoid that the test candidates are distracted by the content of the video sequences as reported by Aldridge et al. [11]. In order to avoid that two consecutive video sequences (e.g., F_2 is following F_1 immediately) have the same content we defined a pattern for the chronological order of the test sessions, as shown in Figure 7. F_x can be any video sequence from the F pool of sequences that has not been used in this specific test session, so far. Thus, a complete test session for a candidate could have a chronological order as shown for Figure 7.

4.4 Layer Patterns

Figure 8 shows the layer patterns of each single sequence that was used in the experiment, except for the first 3 warm-up tests where the comparison is performed between the first sequence that consists of 4 layers and the second that consists of only one layer. Each of the 3 groups shows the patterns that were used with one type of content. Comparisons were always performed between patterns that are shown in a row (e.g., (a1) and (a2)). As already mentioned in Section 1.2 it was our goal to examine fundamental assumptions about the influence of layer

[4] Only watching the sequences and voting took less time, but the candidates had as much time as they wanted to read the questions and possible answers for each test ahead of each test.



Pattern	I_1	I_2	I_3	F_x	M_x	T_x	F_x	M_x	T_x	F_x	M_x	T_x	F_x	M_x	T_x
Sequence 1	I_1	I_2	I_3	F_3	M_1	T_4	F_1	M_2	T_1	F_4	M_3	T_2	F_2	M_4	T_3
Sequence 2	I_1	I_2	I_3	F_3	M_1	T_4	F_1	M_2	T_1	F_4	M_3	T_2	F_2	M_4	T_3

F = Farm
M = Mobile & Calendar
T = Table Tennis

Figure 7. Random generation of test sequence order

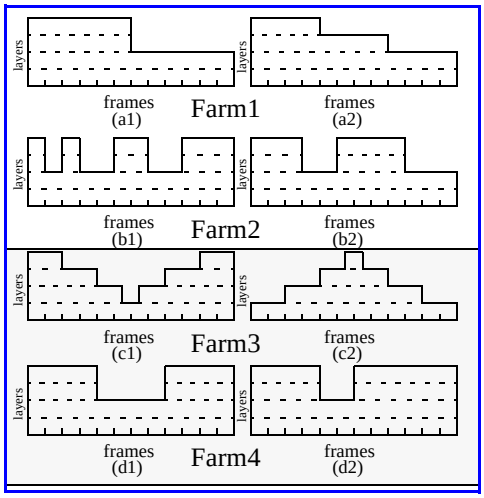
changes on perceived quality. This is also reflected by the kind of patterns we decided to use in the experiment. It must be mentioned that the single layers are not equal in size (contrary to the presentation in Section 8). The size of the n^{th} layer is rather determined by the following expression: $s_n = 2s_{n-1}$. Thus, segments of different layers have different sizes. Preliminary experiments have shown that equal layer sizes are not appropriate to make layer changes perceivable. Since there exist layered schemes that produce layers with sizes similar to ours [25, 26], we regard this a realistic assumption.

In the experiment, we differentiate between two groups of tests, i.e., one group in which the amount of segments used by a pair of sequences is equal and one in which the amount differs (the latter has a shaded background in Figure 8). We made this distinction because we are mainly interested in how the result of this experiment could be used to improve the retransmission scheduling technique (see Section 2.1) where it is necessary to compare the influence of additional segments that is added on different locations in a sequence.

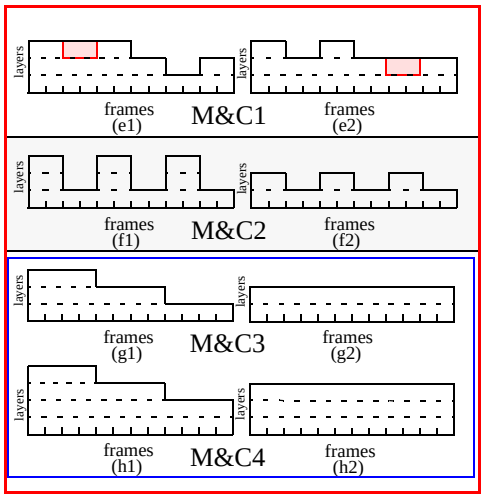
Since segments from different layers are not equal in size, the amount of data for the compared sequences differs. However, somewhat surprisingly, as we discuss in Section 5.3, a larger amount of data does not necessarily lead to a better perceived quality. Additional tests with different quantities of segments in between a pair were chosen to answer additional questions and make the experiment more consistent as we show in Section 5.2.

5. RESULTS

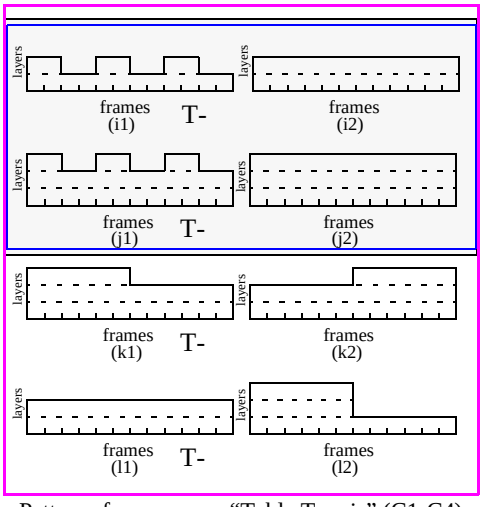
In this section, we present the results of the experiment described in Section 4. Since we analyze the gathered data statistically it must clearly be mentioned that the presented results cannot prove an assumption but only make it less or



Patterns for Sequence “Farm” (F1-F4)



Patterns for sequence “Mobile & Calendar” (M1-M4)



Patterns for sequence “Table Tennis” (C1-C4)

Figure 8. Segments that were compared in the experiment

more likely based on the gathered data. The overall results of all experiments are summarized in Figure 9 and are discussed in the following subsections.

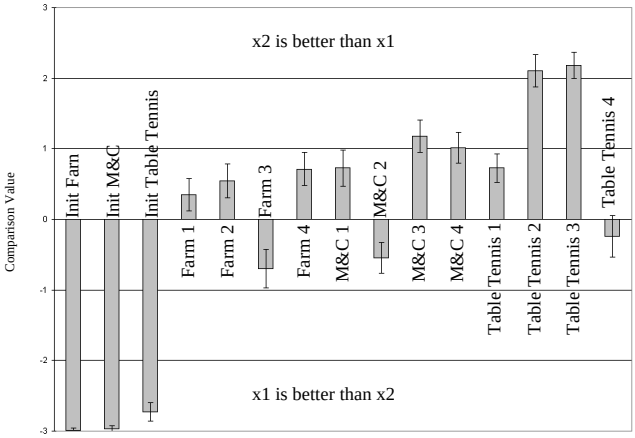


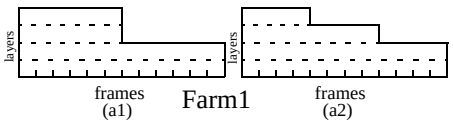
Figure 9. Average and 95% confidence interval for the different tests of the experiment

5.1 Same Amount of Segments

In this section, we discuss the results for the assessments of tests in which the total sum of segments is equal. That means the space covered by the pattern of both sequences is identical.

5.1.1 Farm1: Amplitude

In this assessment the stepwise decrease was rated slightly better than one single but higher decrease. The result shows a tendency that the assumptions that were made about the amplitude of a layer change (as described in Section 2.1) are correct.



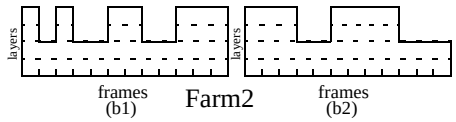
Average:0.35 Conf.-Interval Size:0.23

Figure 10. Farm1

5.1.2 Farm2: Frequency

The result of this test has an even higher likelihood that the second sequence has a better perceived quality than it is the case for *Farm1*. It tends to confirm the assumption that the frequency of layer changes influences the perceived quality,

since, on average, test candidates ranked the quality of the sequence with lesser layer changes better.

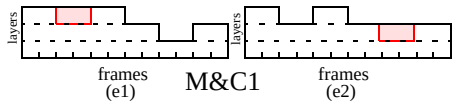


Average:0.55 Conf.-Interval Size:0.24

Figure 11. Farm2

5.1.3 M&C1: Closing the gap

This test should try to answer the question, if it would be better to close a gap in a layer on a higher or lower level. The majority of the test candidates decided that filling the gap on a lower level results in a better quality than otherwise. This result tends to affirm our assumptions made for retransmission scheduling in [7].

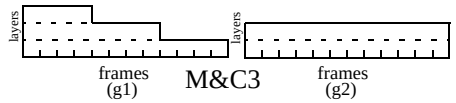


Average:0.73 Conf.-Interval Size:0.25

Figure 12. M&C1

5.1.4 M&C3: Constancy

Even more significant than in the preceding tests, the candidates favored the sequence with no layer changes as the one with the better quality. One may judge this a trivial and unnecessary test, but from our point of view the result is not that obvious, since (g1) starts with a higher amount of layers. The outcome of this test implies that it might be better, in terms of perceived quality, to transmit less but a constant amount of layers.

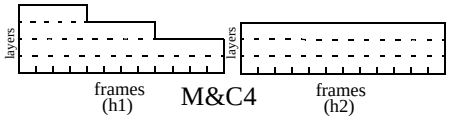


Average:1.18 Conf.-Interval Size:0.23

Figure 13. M&C3

5.1.5 M&C4: Constancy at a higher level

This test was to examine if an increase of the overall level (in this case by comparison to Section 5.1.4) has an influence on the perceived quality. Comparing the results of both tests (*M&C3* and *M&C4*) shows no significant change in the test candidates' assessment. 66% of the test candidates judge the second sequences ((g2) and (h2)) better (values 1-3 in Table 1) in both cases which makes it likely that the overall level has no influence on the perceived quality.

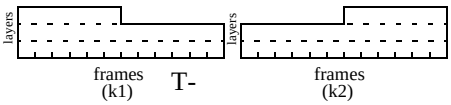


Average:1.02 Conf.-Interval Size:0.22

Figure 14. M&C4

5.1.6 Tennis3: All is well that ends well

The result of this test shows the tendency that increasing the amount of layers in the end leads to a higher perceived quality. The result is remarkably strong (the highest bias of all tests). Future tests, that will be of longer duration and executed in a different order (first (k2) than (k1)), will show how the memory-effect [11] of the candidates influenced this test.

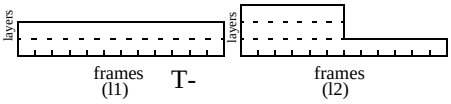


Average:2.18 Conf.-Interval Size:0.19

Figure 15. Tennis3

5.1.7 Tennis4: The exception proves the rule

This test is the only one out of the 12 tests in which the 95% confidence interval covers both areas (better, worth) of the judgement scale. If we regard the average only, the result is a little bit surprising since it contradicts the results from Section 5.1.1 and Section 5.1.4, respectively. At this stage of the investigation, we can only assume that also the content might have an influence on the perceived quality. But to gain more insight in this phenomenon further experiments are necessary.



Average:-0.24 Conf.-Interval Size:0.29

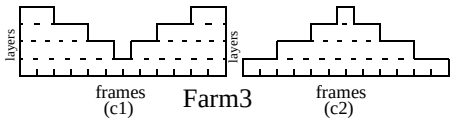
Figure 16. Tennis4

5.2 Different amount of Segments

In the following 5 tests the total amount of segments per sequence differs. All 5 tests have in common that the perceived quality of the sequence consisting of a pattern that covers a larger number of segments were ranked better.
This is obvious, but it makes the overall result more consistent, because test candidates mostly realized this quality difference.

5.2.1 Farm3: Decrease vs. increase

Starting with a higher amount of layers, decreasing the amount of layers, and increasing the amount of layers in the end again seems to provide a better perceivable quality than starting with a low amount of layers, increasing this amount of layers, and going back to a low amount of layers at the end of the sequence. This might be caused by the fact that test candidates are very concentrated in the beginning and the end of the sequence and that, in the first case details become clear right in the beginning of the sequence.

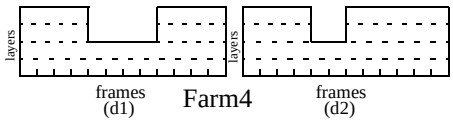


Average:-0.70 Conf.-Interval Size:0.27

Figure 17. Farm3

5.2.2 Farm4: Keep the gap small

In this test, it was our goal to investigate how the size of a gap may influence the perceived quality. The majority of test candidates (37 out of 66) judged the quality of the sequence with a smaller gap slightly better (Only 5 out of 66 judged the first sequence better). This indicates that filling a gap partly can be beneficial.

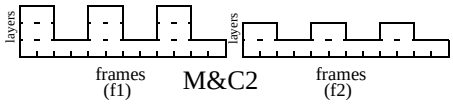


Average:0.71 Conf.-Interval Size:0.23

Figure 18. Farm4

5.2.3 M&C2: Increasing the amplitude

The effect of the amplitude height should be investigated in this test. The result shows that, in contrast to existing assumptions (see Section 2.1), an increased amplitude can lead to a better perceived quality.

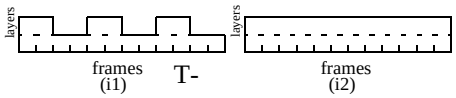


Average:-0.55 Conf.-Interval Size:0.22

Figure 19. M&C2

5.2.4 Tennis1: Closing all gaps

This test is contrary to *M&C2* where the additional segments are used to close the existing gaps instead of increasing the amplitude of already better parts of the sequence. This strategy decreases the frequency of layer changes. Test candidates, on average, judged the sequence without layer changes better. The result of this test reaffirms the tendency that was already noticed in Section 5.1.2, that the perceived quality is influenced by the frequency of layer changes. If we carefully compare the results of *M&C2* and *Tennis1*, a tendency towards filling the gaps and thus decreasing the frequency instead of increasing the amount of already increased parts of the sequence is recognizable. Definitely, further investigations are necessary to confirm this tendency, because, here, the results of tests with different contents are compared and we have not investigated the influence of the content on the perceived quality, so far.

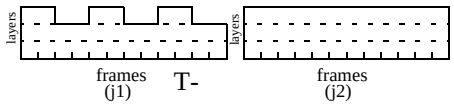


Average:0.73 Conf.-Interval Size:0.20

Figure 20. Tennis1

5.2.5 Tennis2: Closing all gaps at a higher level

In comparison to *Tennis1*, here, we were interested in how an overall increase of the layers (in this case by one layer) would influence the test candidates judgement. Again the sequence with no layer changes is judged better, even with a higher significance than for *Tennis1*. This might be caused by the fact that the amount of layer is higher in general in *Tennis2*.



Average:2.11 Conf.-Interval Size:0.23

Figure 21. Tennis1

5.3 Sequence Size and Quality

As already mentioned in Section 4.4, segments of different layers are not equal in size. Hence, the data size for patterns with an equal amount of segments might not be identical. Here we give an example that shows that sequences consisting of a higher amount of data are not necessarily judged better concerning the perceived quality of the sequence. If we compare, e.g., the tests *M&C1, M&C3,* and *M&C4,* respectively, we recognize that the storage size of the first sequence is always larger than the one of the second. This relation is shown in Figure 22. The results of our experiment show that the average judgement is in contrast to the storage size (Figure 8) where the second sequence has, according to the test candidates, a better perceived quality.

6. CONCLUSION

In this paper, we presented the results of an empirical experiment based on subjective assessment of variations in layer encoded video. A statistical analysis of the experiment

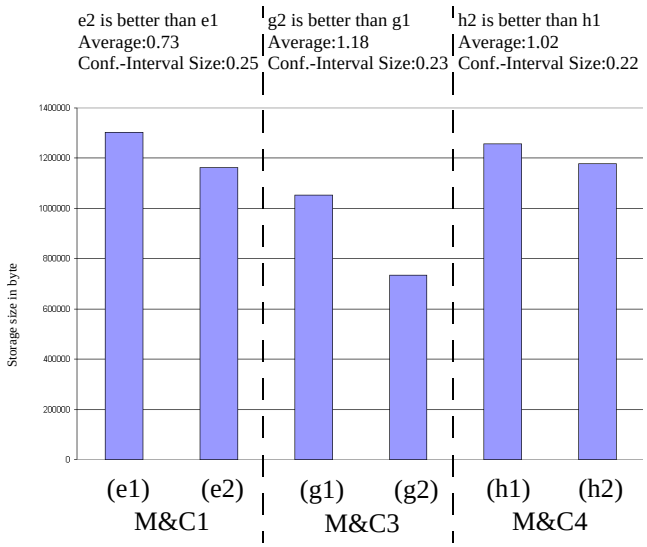


Figure 22. Storage size of sequences

mostly validates assumptions that were made in relation to layer variations and the perceived quality of a video:

- The frequency of variations should be kept as small as possible.
- If a variation can not be avoided the amplitude of the variation should be kept as small as possible.

One basic conclusion from the results in Section 5.2 is: adding information to a layered video increases its average quality. But, as we already assumed in our work on retransmission scheduling, adding information at different locations can have a substantial effect on the perceived quality. Assumptions we made for our heuristics in retransmission scheduling (as well as others' assumptions) could be substantiated by this investigation (see Section 5.1). That means, it is more likely that the perceived quality of a layer encoded video is improved if

- the lowest quality level is increased, and
- gaps in lower layers are filled.

The results from Section 5.3 should be used to refine the retransmission scheduling heuristics in relation to the size of each single layer. Therefore, the metric that represents the quality improvement must also take into account that it might be more expensive to retransmit a segment of layer n+1 than of layer n. Another interesting outcome of the experiment is the fact that a quality improvement may be achieved by retransmitting less data (Section 5.1.3), if a layered encoding scheme is used in which the layers are not of identical size. The obtained results can, in addition, be used to refine caching replacement policies that operate on a layer level [5] as well as layered multicast transmission schemes which try to offer heterogeneous services to different subscribers as, e.g., in the receiver-driven layered multicast RLM [27] scheme and its derivations.

The results of this investigation clearly strengthen the assumption that a differentiation between objective and subjective quality, in the case of variations in layer encoded video, must be made.

Nevertheless, it must be admitted that the presented work is only an initial investigation in the subjective impression of variations in layer encoded videos. In further work, we want to explore sequences with a longer duration (up to several minutes). In a next step, we will investigate if the shown sequences can be combined and if the subjective assessment is still consistent with the separated results. E.g., in this experiment sequences (e2) and (g2) were judged better than (e1) and (g1), will a sequence that combines (e2) and (g2) also be judged better than a sequence that combines (e1) and (g1)? We are also interested in how the content of a sequence influences the perceived quality.

7. ACKNOWLEDGMENTS

The authors would like to thank Rico Tunk for creating the test application, Charles "Buck" Krasic for his support on SPEG, the test candidates for taking the time to perform the assessment, and RTL Television for providing the video sequences.

8. REFERENCES

[1] J. Lu. Signal Processing for Internet Video Streaming: A Review. In *Proceedings of SPIE Image and Video Communications and Processing, San Jose, CA, USA*. SPIE - The International Society for Optical Engineering, January 2000.

[2] D. Saparilla and K. W. Ross. Optimal Streaming of Layered Video. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies 2000 (INFOCOM'00), Tel-Aviv, Israel*, pages 737–746, March 2000.

[3] S. Nelakuditi, R. R. Harinath, E. Kusmierek, and Z.-L. Zhang. Providing Smoother Quality Layered Video Stream. In *Proceedings of the 10th International Workshop on Network and Operating System Support for Digital Audio and Video, Raleigh, NC, USA*, June 2000.

[4] R. Rejaie, M. Handley, and D. Estrin. Quality Adaptation for Congestion Controlled Video Playback over the Internet. In *Proceedings of the ACM SIGCOMM '99 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication 1999, New York, NY, USA*, pages 189–200, August 1999.

[5] R. Rejaie, H. Yu, M. Handley, and D. Estrin. Multimedia Proxy Caching for Quality Adaptive Streaming Applications in the Internet. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies 2000 (INFOCOM'00), Tel-Aviv, Israel*, pages 980–989, March 2000.

[6] R. Rejaie and J. Kangasharju. Mocha: A Quality Adaptive Multimedia Proxy Cache for Internet Streaming. In *Proceedings of the 11th International Workshop on Network and Operating System Support for Digital Audio and Video, Port Jefferson, New York, USA*, pages 3–10, June 2001.

[7] M. Zink, J. Schmitt, and R. Steinmetz. Retransmission Scheduling in Layered Video Caches, April 2002. Accepted at ICC 2002, New York, New York, USA.

[8] T. Alpert and J.-P. Evain. Subjective quality evaluation - The SSCQE and DSCQE methodologies. EBU Technical Review, February 1997.

[9] E. Kohler, M. Handley, S. Floyd, and J. Padhye. Datagram Control Protocol (DCP). Internet Draft, November 2001. Work in Progress.

[10] ITU-R: Methodology for the Subjective Assessment of the Quality of Television Picture. International Standard, 2000. ITU-R BT.500-10.

[11] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson. Measurement of scene-dependent quality variations in digitally coded television pictures. *IEE Proceedings on Vision, Image and Signal Processing*, 142, 3:149–154, 1995.

[12] R. Aldridge, D. Hands, D. Pearson, and N. Lodge. Continuous quality assessment of digitally-coded television pictures. *IEE Proceedings on Vision, Image and Signal Processing*, 145, 2:116–123, 1998.

[13] F. Pereira and T. Alpert. MPEG-4 video subjective test procedures and results. *IEEE Transactions on Circuits and Systems for Video Technology*, 7, 1:32–51, 1997.

[14] M. Masry and S. Hemami. An analysis of subjective quality in low bit rate video. In *International Conference on Image Processing (ICIP), 2001, Thessaloniki, Greece*, pages 465–468. IEEE Computer Society Press, October 2001.

[15] C. Kuhmünch and C. Schremmer. Empirical Evaluation of Layered Video Coding Schemes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece*, pages 1013–1016, October 2001.

[16] T. Hayashi, S.Yamasaki, N. Morita, H. Aida, M. Takeichi, and N. Doi. Effects of IP packet loss and picture frame reduction on MPEG1 subjective quality. In *3rd Workshop on Multimedia Signal Processing, Copenhagen, Denmark*, pages 515–520. IEEE Computer Society Press, September 1999.

[17] S. Gringeri, R. Egorov, K. Shuaib, A. Lewis, and B. Basch. Robust compression and transmission of MPEG-4 video. In *Proceedings of the ACM Multimedia Conference 1999, Orlando, Florida, USA*, pages 113–120, October 1999.

[18] M. Chen. Design of a virtual auditorium. In *Proceedings of the ACM Multimedia Conference 2001, Ottawa, Canada*, pages 19–28, September 2001.

[19] S. Lavington, N. Dewhurst, and M. Ghanbari. The Performance of Layered Video over an IP Network. *Signal Processing: Image Communication, Elsevier Science*, 16, 8:785–794, 2001.

[20] C. Krasic and J. Walpole. Priority-Progress Streaming for Quality-Adaptive Multimedia. In *ACM Multimedia Doctoral Symposium, Ottawa, Canada*, October 2001.

[21] C. Krasic and J. Walpole. QoS Scalability for Streamed Media Delivery. Technical Report OGI CSE Technical Report CSE-99-011, Oregon Graduate Institute of Science & Technology, September 1999.

[22] Intel. Developers - What Intel Streaming Web Video Software Can Do For You, 2000. http://developer.intel.com/ial/swv/developer.htm.

[23] PacketVideo. Technical White Paper: PacketVideo Multimedia Technology Overview, 2001. http://www.packetvideo.com/pdf/pv_whitepaper.pdf.

[24] R. Neff and A. Zakhor. Matching Pursuit Video Coding–Part I: Dictionary Approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 12, 1:13–26, 2002.

[25] J. Hartung, A. Jacquin, J. Pawlyk, and K. Shipley. A Real-time Scalable Software Video Codec for Collaborative Applications over Packet Networks. In *Proceedings of the ACM Multimedia Conference 1998, Britol, UK*, pages 419–426, September 1998.

[26] L. Vicisano, L. Rizzo, and J. Crowcroft. TCP-like congestion control for layered multicast data transfer. In *Proceedings of the 17th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'98)*, pages 996–1003. IEEE Computer Society Press, March 1998.

[27] S. McCanne, M. Vetterli, and V. Jacobson. Receiver-driven layered multicast. In *Proceedings of ACM SIGCOMM'96*, Palo Alto, CA, August 1996.