Modelling Video-Quality Shaping with Interpolation and Frame-Drop Patterns

Håvard Berge, Mauritz Panggabean[1], and Leif Arne Rønningen[1]

[1] Department of Telematics

Norwegian University of Science and Technology (NTNU)

O.S. Bragstads plass 2B, N-7491, Trondheim, Norway

{panggabean,leifarne}@item.ntnu.no

Abstract

This paper investigates combining video frames of different spatial resolutions to model and study sub-object dropping based on the concept of Quality Shaping in the Distributed Multimedia Plays (DMP) architecture for future networked collaboration. It is aimed at reducing the data rate needed to stream video data over a network while keeping the perceived video quality as high as possible. Frame-drop patterns are introduced for this purpose and two different patterns were examined using two different temporal resolutions and using different frame-drop patterns and spatial resolution on the low quality frames. The results are assessed in a qualitative manner with an action-research approach. The experimental results show that different frame-drop patterns and spatial resolution on low quality frames can be combined to have a large impact on the perceived quality of the resulting video sequence. Some combinations can give a perceived quality almost as good as the original one without any frame drop or resolution reduction, saving network bandwidth considerably.

1 Introduction

Rapid advancements in electronics as well as information and communication technology (ICT) have unveiled the possibility of new and more creative ways of real-time multiparty collaborations limited only by time and space. We envision such collaborations in the future with near-natural quality of experience through networked collaboration spaces that seamlessly combine virtual (taped) and live scenes from distributed sites on the continents that may have different technical specifications to each other. For example, as shown in Figure 1, an audience in Oslo (A) are attending a concert from two opera singers in a specially designed room, namely a collaboration space. The multimedia quality that they experience is so close to natural that they hardly realize that two singers are singing live from two different cities, say Trondheim (B) and Tromsø (C), each in their own collaboration space. The two singers perform together so harmoniously with life-like multimedia quality that the audience feel they are enjoying a live opera concert

This paper was presented at the NIK-2010 conference; see http://www.nik.no/.

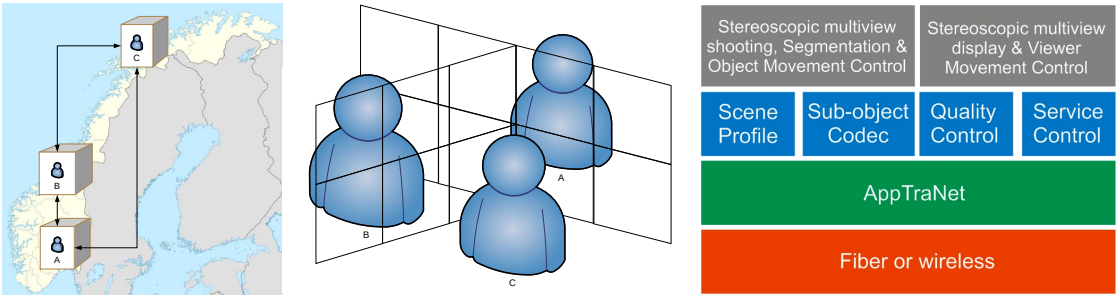

Figure 1: A simple example of the envisioned collaboration (left) and the corresponding combined collaboration spaces (middle). Arrays of multiview 3D display, dynamic cameras, speakers and microphones are proposed for all the surfaces of the collaboration space. The proposed three-layer DMP architecture from a user's perspective (right).

Table 1: The main technical requirements for the envisioned collaborations derived from the aimed quality of experience.

Nr.	Main technical requirements
1.	Guaranteed maximum end-to-end delay $\leq$ 10-20ms
2.	Near-natural video quality
3.	Autostereoscopic multi-view 3D vision
4.	High spatial and temporal resolution due to life-size dimension of objects i.e. mostly humans
5.	Accurate representation of physical presence cues e.g. eye contact and gesture
6.	Real 3D sound
7.	Quality allowed to vary with time and network load due to different technical specifications and behaviors among collaboration spaces
8.	Minimum quality guarantee using admission control
9.	Graceful quality degradation due to traffic overload or failure
10.	Privacy provided by defined security level

and they are in the very same room with the two singers. Moreover, each opera singer performing live also experiences singing together with the other two displayed in his or her own collaboration space, as if they are on the same stage.

The main technical requirements on important aspects for the envisioned collaborations are listed in Table 1. The ultimate challenge is ensuring that the maximum end-to-end delay is less than 20 ms or even 10 ms to enable good synchronization in musical collaboration. Chafe et al. [1] reports experimental results on the effect of time delay on ensemble accuracy by placing pairs of musicians apart in isolated rooms and asking them to clap a rhythm together. It shows that longer delays produced increasingly severe tempo deceleration and shorter delays produced a modest, but surprising acceleration. The result indicates that the observed optimal delay for synchronization is 11.5 ms that equates with a physical radius of 2,400 km (assuming signals traveling at approximately 70% the speed of light and no routing delays). Realizing such collaborations with very high quality and also complexity will only be possible if the rest of the requirements can be fulfilled within the maximum time delay.

As current standards are still unable to realize this vision, we propose the three-layer

Distributed Multimedia Plays (DMP) architecture [2], as illustrated by the image on the left in Figure 1. In our proposal, we also introduce the concept of Quality Shaping that enables graceful degradation of quality when traffic overloads the network or system components fail. A more detailed presentation of the concept is provided in Section 2.

Traffic generated from near-natural scenes is extremely high, up to 10^3 – 10^4 higher than from today's videoconferencing systems. 'Near-natural' quality requires an extreme resolution, which means data rates of Gbps, even for a human face alone. Moreover this traffic is also extremely variable during the collaboration. These make simulation of such traffic extremely difficult. In this work, we approach the study of such traffic and the perceived video quality by modelling the Quality Shaping with interpolation and frame-drop patterns, as elaborated in Section 2.

In the context of the envisioned collaboration and the modelled Quality Shaping, the objective of this work is two-fold. First, to investigate the perceived quality of moving and static objects in a video sequence when time-variable quality is applied on short time intervals, and second, to identify parameters with their combinations that affect the perceived quality considerably when time-variable quality is used.

The organization of the paper is as follows. The objective is addressed by the two experiments explained in Section 3 which follows the concept of Quality Shaping and its modelling briefly presented in Section 2. Section 4 presents the experimental results with the evaluations that lead to our conclusions in Section 5 as our main contributions.

2 Quality Shaping, Interpolation and Frame-Drop Patterns

Quality Shaping [2] is a quality control scheme used in DMP networks that allows the quality of scenes to vary with the network load, but guarantees that the quality will not be lower than a predefined minimum value. DMP network nodes implement dropping of sub-object packets and also controlled feedback of measured traffic parameters to traffic sources. See [2] fore more details. This allows end users to enjoy good experience of viewing the scenes despite necessary quality reduction. Video scenes are composed of real and virtual (stored or generated) sub-scenes while sub-scenes consist of hierarchies of sub-scenes. The smallest real entity in a sub-scene is called object and an object can be divided further into sub-objects, as illustrated in Figure 2. This is the basis for making multimedia content packets independent and utilizing independent parallel computations, which is the goal of the concept to be able to meet the maximum end-to-end delay. There has been much research with many proposed techniques for object segmentation for different applications. The concept of Quality Shaping builds upon the concept of Traffic Shaping first proposed in [3] which evaluated variance reduction of traffic streams using queues and controlled dropping of packets. Therefore, the model of the analyzed sequences in this work focuses on objects, sub-objects in video scenes and their quality which results from changing the combinations of three main parameters, i.e. the frame-drop pattern, the spatial resolution, and the frame rate.

One simple way to model sub-objects is by allocating a set of numbers to pixel in rectangular form. An $N \times N$ square of pixels will contain N^2 sub-objects. The ith sub-object in such a square is denoted by S_i where $i = 1, 2, 3, ..., N^2$. Each sub-object will be processed and transmitted independently as independent bit streams. The number of sub-objects and which sub-objects to be dropped in the transmission must be handled properly such that graceful degradation of the video quality can be achieved to meet the maximum end-to-end delay due to network condition.

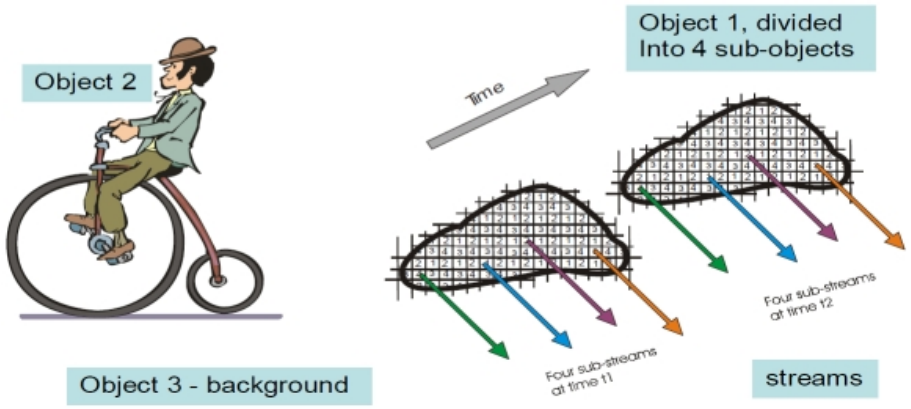

Figure 2: An example of object-oriented scene with sub-objects in a video stream.

1	2	3	1	2	3
4	5	6	4	5	6
7	8	9	7	8	9
1	2	3	1	2	3
4	5	6	4	5	6
7	8	9	7	8	9

1	2	x	1	2	x
x	5	6	x	5	6
7	x	9	7	x	9
1	2	x	1	2	x
x	5	6	x	5	6
7	x	9	7	x	9

Figure 3: Object and nine sub-objects (left); sub-object 3, 4, and 8 are dropped (right).

A	B
C	D

A			B
C			D

1	2	3	1
4	5	6	4
7	8	9	7
1	2	3	1

Figure 4: Received original pixels (left) interpolated with twice magnification (middle) compared to a 3×3 square with 9 sub-objects (right).

Figure 3 includes 3×3 squares each with 9 sub-objects and the left image depicts the dropped sub-objects 3, 4 and 8. Figure 4 illustrates the problem of interpolation with twice magnification which is identical to the tiling of 3×3 squares where only sub-object 1 is retained and the other eight are dropped. Dropping 8 out of 9 sub-objects is the most extreme case in Quality Shaping that must be very rare. Quality Shaping and interpolation have the same goal, which is to estimate the values of the dropped sub-objects from the retained ones (for Quality Shaping) or to estimate the values of the interpolated pixels from the original ones (for interpolation). Thus we argue that Quality Shaping can be modeled with interpolation, although their initial conditions are not the same. There are some well-known interpolation techniques for two-dimensional data and digital images, for example, nearest neighborhood, bilinear and bicubic interpolations.

The objective of all interpolation algorithms is to estimate the value $f_0 = f(x,y)$ at a given location (x_0, y_0) as a weighted sum of data values at neighboring S locations, formulated as

$$f(x,y) = \sum_{s=1}^{S} w_s f(x_s, y_s)$$

where S denotes the number of the known sample values of the signal and w_s is the corresponding weight. Almost all interpolation techniques use functions that give a

decreasing weight to samples values at further distance. Let us introduce *Kriging* as a more suitable interpolation technique for the arbitrary dropping of sub-objects in Quality Shaping as illustrated in Figure 3. Kriging has been widely used in fields related to geology, but very rarely applied in image processing, particularly for image super-resolution such as in [4]. Detailed treatise on Kriging, particularly from the perspective of geostatistics, can be found in [5, 6] that base the following brief explanation on the basics of Kriging as a formal approximate model of sub-object dropping in Quality Shaping.

Kriging considers both the distance and the degree of variation between data points when estimating the values in unknown areas. Different than other interpolation techniques, Kriging attempts to minimize the error variance and set the mean of the prediction errors to zero. Hence Kriging helps to compensate for the effects of data clustering by assigning individual points within a cluster less weigh then isolated data points. The basic form of the Kriging estimator is expressed by

$$Z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha [Z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha)]$$

and the goal of Kriging is to determine the weights λ_α that minimize the variance of the estimator

$$\sigma_E^2(\mathbf{u}) = Var\{Z^*(\mathbf{u}) - Z(\mathbf{u})\}$$

under the unbiasedness constraint $E\{Z^*(\mathbf{u}) - Z(\mathbf{u})\} = 0$. The random variable $Z(\mathbf{u})$ is decomposed into residual and trend components $Z(\mathbf{u}) = R(\mathbf{u}) + m(\mathbf{u})$ with the residual component $R(\mathbf{u})$ treated as a random variable with zero stationary mean $ER(\mathbf{u}) = 0$ and a stationary covariance as a function of lag $\mathbf{h}$ and not of position $\mathbf{u}$ as follows

$$Cov\{R(\mathbf{u}), R(\mathbf{u}+\mathbf{h})\} = E\{R(\mathbf{u}).R(\mathbf{u}+\mathbf{h})\} = C_R(\mathbf{h}).$$

The residual covariance function is generally derived from the input semivariogram model that should represent the residual component of the variable. The shape of the semivariogram can be approximated by a mathematical model, such as a Gaussian model. As almost all interpolation techniques, however, Kriging will give very similar results by

- producing quite good estimates given fairly dense uniformly distributed data locations throughout the study area,
- yielding unreliable estimates if the data locations fall in a few clusters with large gaps in between, and
- overestimating the lows and underestimating the highs, as inherent to averaging.

However Kriging is still under preliminary investigation in our research and therefore this work is still based on the classic interpolation techniques used in image processing, particularly the bicubic interpolation.

Ideally the quality reduction in a scene is applied to its objects and sub-objects after intelligent object segmentation, which has been an active research topic. Our approach in this work is to apply the quality reduction generally to the whole scene by reducing the frame resolution, by downsampling and regeneration of the missing pixels by interpolation. In this paper, we introduce the concept of *frame-drop pattern* with its simple notation which is a model of sub-object dropping that indicates how the frames in a video

sequence are structured whether as a frame with high spatial resolution (denoted by '1') or with reduced spatial resolution (denoted by '2') after interpolation. For example, frame-drop pattern 1:2 denotes a cyclic pattern where one original quality frame is followed by one frame with reduced spatial resolution. Frame-drop pattern 1:1 represents a sequence with all frames in original high resolution. Quality shaping is thus represented by the frame-drop pattern with interpolation.

This work continues some of the work reported in [7] and [8]. The first addresses the quality variation of video sequences with intervals of 8-10 seconds. It concludes that regeneration of dropped sub-objects using linear interpolation in many cases is very good for regaining visual quality. This work also used interpolation for low quality frames with reduced spatial resolution. It also showed that applying edge correction in most cases can further increase the perceived quality. In real networks the interval will vary randomly from a few milliseconds to several seconds depending on characteristics and load of the traffic, this work will focus on perception of video sequences with quality variation in interval within tenths of milliseconds.

3 Method and Setup of the Experiments

As there are ways to vary quality of the objects over time, every assessor would have to view a number of different video sequences and focus on the objects with necessary guidance from us. Therefore, we argue that it is best for preliminary research such as this one to focus on a relatively small number of assessors and spend more time to really understand how they perceive the assessed sequences because we emphasize on certain objects in the scenes and their quality. The experiments were thus conducted and the results were assessed in a qualitative manner with action-research approach as qualitative research requires more involvement from the researcher to spend more time with each research subject [9]. Given a limited time frame, qualitative research allows the use of less research subjects than quantitative research. Following the perspective of action research, there were interviews with each assessor to improve how the next assessment is performed and to find if anything is unclear or difficult to understand. Cycles of active guidance to each assessor during the assessment process allow better description of what they should look for and any irregularities or aspects of the video sequences that they can ignore.

The first and second experiments in this work followed the single-stimulus adjectival categorical judgment method (SSACS) and the stimulus-comparison adjectival categorical judgment method (SCACS) [11], respectively, with the following adaptations. First, the duration of the video sequences in the experiments (21 seconds) is longer than recommended (10 seconds). Second, as there is only one 46" HDTV available for viewing and assessment, both the reference and the assessed sequences had to be shown on the same display device. However, when objects are moving fast, it is difficult to follow the objects on two displays simultaneously. Third, it is recommended that the participants are not experts or experienced in assessing video quality, but for a preliminary research such this one, it is beneficial that participants have some experience to be able to rule out some error sources. Fourth, for more accurate assessment, ten values instead of the five recommended by the ITU-R were used in Experiment-1. Moreover, decimal numerals were allowed in Experiment-2 instead of values from -3 to +3.

In all video sequences in both experiments, a spatial true HD resolution of 1920×1080 and two temporal resolutions (29.97 fps and 59.94 fps) were used as they are the highest spatial and temporal resolution of public TV broadcasts today, respectively. This choice

Figure 5: A scene in the reference sequence in this work showing background, static objects and dynamic objects to be assessed with altered frame-drop patterns as well as spatial and temporal resolutions.

is driven by lack of more advanced equipment. The sequences were displayed on a 46" 120Hz HDTV monitor supporting spatial resolutions up to 1920×1080 with 60 fps temporal resolution. Following the concept of Quality Shaping, the model of the reference and modified sequences had a background, static objects and dynamic objects and each will be assessed in the experiments. Figure 5 depicts a scaled scene from the reference sequence in this work. The two cassette boxes were mounted on a controllable moving conveyor belt and they function as the dynamic objects. The wall and the other static objects in the scene serve as the background and the static objects, respectively.

The scene was recorded as uncompressed AVI video with such settings that the playback on the HDTV and the factual recorded model are identical in size. Figure 5 also shows that the lighting was set such that shadows were minimized and objects were clearly visible. According to [10], for a person with perfect 20/20 vision, the best pixel size is the distance of the viewer from the TV divided by 3400. With 1.06 meters as the horizontal side of the HDTV, the best distance of the assessor to the TV is then 1.88 meters. Five assessors were involved in the two experiments and all have a certain degree of knowledge about video quality.

Experiment-1

The objective of Experiment-1 is to combine frames of low and high resolution according to predefined frame-drop patterns to find out the effects of each combination to the perceived quality for different objects in the scene. Each assessor sat in front of the HDTV and the reference sequence was shown. Following the action-research approach, some of the scene characteristics in the sequence and parts considered as background, static objects and dynamic objects were explained to the assessor. The assessors are educated on what to look for when assessing the sequences.

After the reference sequence had been shown and the assessor understood the expected qualities, the corresponding sequence to be assessed was displayed. We asked the assessor to first look for specific qualities and variation in the quality of the background, the static objects and the dynamic object, and in the end a general impression of the assessed sequence. Then we filled in the questionnaire with the information from the assessor and

Table 2: Frame-drop patterns and frame rates for video sequences in Experiment-1

Seq. nr.	Frame-drop pattern	Frame rate (fps)
1	1:1:2:2	29.97
2	1:1:2:2	59.94
3	1:2:1:2	29.97
4	1:2:1:2	59.94

Table 3: Frame-drop patterns and resolutions of the low quality frames for video sequences in Experiment-2

Seq. nr.	Frame-drop pattern	Resolution
1	1:1:1:2	960×540
2	1:1:2	1024×576
3	1:2:2	1440×810
4	1:2:2:2	1600×900

may use it to make necessary improvements in the next assessment for the next assessor.

The video sequences assessed in this experiment used two different frame-drop patterns (1:1:2:2 and 1:2:1:2) and temporal resolutions of 29.97 and 59.94 fps, as shown in Table 2. This experiment was conducted according the single-stimulus adjectival categorical judgment method (SSACS) with some adaptations, as detailed in [8].

Experiment-2

Experiment-2 was conducted similarly to Experiment-1. The main differences between them are the method of the assessment on semantic terms and the assessment scale used. The reference video sequence used in Experiment-2 was the video sequence with frame-drop pattern similar to that of the sequence with the best result in Experiment-1.

This experiment investigates whether a higher bit rate always give a better viewing experience and examines various combinations of different resolutions and frame-drop patterns to achieve perceived quality that is as good as that of the video with a higher bit rate. The temporal resolution was maintained at 59.94 fps for all video sequences, while the resolution of the low quality frames and the frame-drop patterns vary, as detailed in Table 3.

4 Experimental Results and Evaluations

This section presents the results of the two experiments that are followed by the evaluations.

Experiment-1

Table 4 presents the assessment result from Experiment-1 with four modified sequences shown in Table 2. Each sequence is evaluated based on the assessment result. Theoretically Sequence-1 has the worst quality and it is confirmed by the lowest number in every category (general, background, static objects, and dynamic objects). Sequence 1

Table 4: Assessment result from Experiment-1.

Nr.	General				Background				Static objects				Dynamic objects			
	Video sequence number															
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1	7	8	8	9	8	8	8	10	7	7	8	9	8	9	8	10
2	3	4	4	7	6	7	7	9	2	2	2	4	4	5	6	7
3	5	6	6	9	6	8	8	10	3	4	4	9	8	10	8	10
4	5	6	5	9	5	6	6	10	4	4	4	9	6	7	6	10
5	5	7	6	8	7	7	7	9	4	6	6	7	5	7	7	9
Avg	**5,0**	**6,2**	**5,8**	**8,4**	**6,4**	**7,2**	**7,2**	**9,6**	**4,0**	**4,6**	**4,8**	**7,6**	**6,2**	**7,6**	**7,0**	**9,2**

had small ”jumps” as it shifted between the high and low resolution frames. These ”jumps” were very clear with the static objects. These caused the quality of the background suffered a little, although better than that of the static objects. The dynamic object was not influenced by the jumping phenomena as clearly as the other parts of the scene. It is most likely because the motion of the dynamic object made the jumps very hard to see. The largest decrease in perceived quality for the dynamic object was shown by the less smooth motion than that in the reference sequence. This is logical as the video sequence used a 29.97 fps temporal resolution compared to the 59.94 fps of the reference.

Sequence-2 was generally better than the first. The issues with the image jumping still exist although much less noticeable due to the doubled frame rate relative to that of Sequence-1. Instead of frames shifting between high and low quality every 66.73ms as in Sequence-1, it shifted every 33.36ms which makes it harder for human eyes to see the changes. The dynamic object had a smoother motion than that in Sequence-1 and one assessor even stated that he could now not see any difference when compared to that in the reference sequence. The most disturbing part of the scene was still the static objects due to the jumps. The jumps were less obvious with the background probably because it does not have the clear contrasts and graphic like those of the static objects. There are fewer details that were clearly observed when the jumps occurred.

Sequence-3 was expected to perform similarly to Sequence-2 as the change between low and high quality frames happened at the same frequency, i.e. every 33.36ms as in Sequence-2. This is confirmed by the values which are close to those of Sequence-2. However, this change of Sequence-4 happened every 16.68ms. Although the jumping in the static object was very hard to see in Sequence-4, some noise are still apparent. Nevertheless, the overall quality of Sequence-4 was very close to that of the reference sequence. This comes with an impressive 37.5% less bit rate than that of the reference based on the actual file sizes and frame rates.

The results show clearly that the perceived quality increases when the duration of the low resolution frames is shorter. The ”jumping” that caused the most perceived quality loss when sequences are compared to the reference is caused by the pixel interpolation. Figure 6 gives an enlarged illustration on the effects of various interpolation techniques from 960×540 resolution (images 2-5) to 1920×1080 resolution (image 1 as the reference). These images of the ’S’ letter on the static object in the scene are zoomed 3200% of the actual resolution. Images 2 to 5 are the result of bilinear interpolation, bicubic interpolation, soft interpolation, and nearest-neighbor interpolation, respectively. The edges of the letter ’S’ in the last four images are shifted a few pixels to the right and upwards as the cause of the jumping phenomenon.

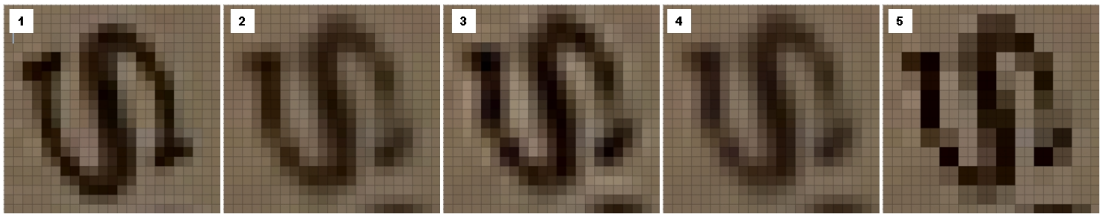

Figure 6: The effects of various interpolation techniques from 960×540 resolution (images 2-5) to 1920×1080 resolution (image 1 as the reference). Images 2 to 5 are the result of bilinear interpolation, bicubic interpolation, soft interpolation, and nearest-neighbor interpolation, respectively.

Table 5: Assessment result from Experiment-2.

Nr.	General				Background				Static objects				Dynamic objects			
	Video sequence number															
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	-2	-1	-1	-0.5	-0.5	-0.2	-0.2	-0.1	-2	-1.5	-1.5	-1	0	0	0	0
2	-2	-1	-1	0	-1	0	1	2	-2	-1	-1	-1	-1	-1	0	0
3	-1	-0.5	-0.5	-0.5	-1	0	0	0	-2	-1	-1	-1	0	1	1	1
4	-3	-1.5	-0.5	-1	0	0	0	0	-3	-2	-1	-2	0	0	0	0
5	-1	-1	-1	-0.5	-1	-0.5	-1	-0.5	-1.5	-1	-0.5	-0.5	0	0	0	0
Avg	**-1.8**	**-1**	**-0.8**	**-0.5**	-0.7	**-0.14**	**-0.04**	**0.28**	**-2.1**	**-1.3**	**-1**	**-1.1**	**-0.2**	**0**	**0.2**	**0.2**

Experiment-2

Table 5 presents the assessment result from Experiment-2 with four modified sequences shown in Table 3. Each sequence is evaluated based on the assessment result. Sequence-1 gave the worst overall perceived quality in Experiment-2. This is surprising as it has the highest bit rate of all the assessed sequences. The most obvious reduction in quality was with the static object suffering from the same jumping happened in Experiment-1. Sequence-2 had less noticeable jumping with the background and the static object, but it was much more obvious than it was with the reference sequence. Although the overall quality was far from that of the reference sequence, it was slightly better than Sequence-1. The jumping in Sequence-3 was about the same as that in Sequence-2. The overall quality was below that of the reference sequence. Details in the dynamic object were clearer to see when the frames did not change between low and high quality so often. Sequence-4 gave the highest overall quality impression, but it was still not as good as the quality of the reference sequence. The quality of the background can be hard to assess as the jumping is not that obvious. The static objects in Sequence-4 got a lower score too due to the jumps.

These result indicates that none of the assessed sequences gave an overall impression of higher quality than that of the reference sequence. This is caused by jumps with the static object which are more apparent when the frame-drop pattern changed from the 1:2:1:2 pattern. This problem is more noticeable whenever the change between high and low quality frames happens in intervals where one type of frame lasts longer than the other, even if the majority of frames are of high quality as was the case for Sequence-1 where 75% of the frames have a 1920x1080 resolution. The jumping is less noticeable when the duration of low and high quality frames are almost equal as with Sequence-2 and 3 which 66.6% of the frames were of either high or low quality. It became more noticeable again for Sequence-4 where 75% of the frames were of low resolution. It is

surprising that for Sequence-4, which had 1600×900 as the low resolution frames, the jumping was so clear. The cause was again that the pixels were shifted up and to the right. Therefore, increasing bit rate of a video sequence does not necessarily increase its perceived quality. How the frames of low and high resolution are combined has a big impact on the resulting perceived quality of a video.

From the results of the experiments, it is very important that the static objects are shown properly since all assessors noticed the biggest degradation in quality in them when resolution and frame-drop patterns changed. For the background part, the assessors mostly commented on a degraded quality when it was jumping because of changes between high and low quality frames. The quality of the dynamic object was generally maintained at different frame-drop patterns and resolutions. This may be caused by its lower quality than that of the static objects. The dynamic object is often an important part in video scenes so its quality should be maintained high.

Taking these into consideration, if the objects in the sequences can be automatically and intelligently segmented, the recommended parameters for the different objects to achieve the objective are as follows:

- Background: 2:2 frame-drop pattern at 1280×720 resolution
- Static objects: 1:2:1:2 frame-drop pattern with 960×540 resolution for low quality frames
- Dynamic object: 1:2:1:2 frame-drop pattern with 1280×720 resolution for low quality frames.

Assuming that the scenes in the sequences generally consist of 75% background, 15% static objects and 10% dynamic objects, based on the actual file sizes, we can expect with the recommended parameters to have up to twice the number of viewers by reducing the quality of objects in the scene while maintaining the overall perceived quality of the streamed video high.

5 Conclusion

The proposed Quality Shaping introduces good possibilities for saving bandwidth in a network and thus allows more users to connect to the network. Kriging interpolation has been introduced and briefly presented as a formal approximate model of the sub-objects dropping in Quality Shaping and also a promising solution to the problem of estimating the values of the sub-objects dropped in a controllable fashion. When few users are simultaneously active, they can obtain very good quality. However, using spatial interpolation to reduce the quality in representing Quality Shaping causes jumping phenomena as shifts in pixel positions. Four different interpolation techniques were used in this work, but none is free from the jumping. This may be solved by applying edge correction after interpolation that can result in substantial improvement in perceived quality [7]. Some frame-drop patterns give much better perceived quality than others. The network traffic behavior may force certain frame-drop patterns, and an oscillation effect can happen when the network cannot stabilize itself. The different frame-drop patterns examined in this work can represent such oscillation patterns.

Experiment-2 showed that a high bit rate does not always produce high perceived video quality. Adaptive and intelligent combinations of resolution and frame-drop

patterns are challenges to addressed in the future as a way to optimize the use of bandwidth. It is possible to save as much as 37.5% of total bandwidth by only alternating high and low resolution frames without causing a noticeable reduction in perceived quality. The possibility of saving bandwidth is even larger if a scene can be automatically segmented into objects and sub-objects as required by DMP. Controlled dropping of packets in the network allows us to reduce the resolution of objects in a scene so the overall size of the media stream is reduced. Even with an increase in network capacity, intelligent use of the available resources will always be necessary. Quality Shaping is needed to enable this in a controlled manner and to guarantee minimal video quality.

References

[1] C. Chafe, M. Gurevich, G. Leslie and S. Tyan. Effect of time delay on ensemble accuracy. In Proc. International Symposium on Musical Acoustics. 2004.

[2] L.A. Rønningen. The DMP system and physical architecture. Technical report, Department of Telematics, Norwegian University of Science and Technology, Trondheim, Norway, 2007.

[3] L.A. Rønningen. Input Traffic Shaping. In Proc. International Teletraffic Congress, Montreal, Canada, 1982.

[4] A. Panagiotopoulou and V. Anastassopoulos. Super-resolution image reconstruction employing Kriging interpolation technique. In Proc. IWSSIP and EC-SIPMCS, 2007.

[5] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, 1997.

[6] E. H. Isaaks and R. M. Srivastava. *Applied Geostatistics*, Oxford University Press, New York, 1989.

[7] L.A. Rønningen and E. Heiberg. Perception of time variable quality of scene objects. In Proc. SPIE Image Quality and System Performance VI, 2009, vol. 7242.

[8] H. Berge. The Hems Lab - Perceptual test of scene objects with variable temporal resolution. Project report, Department of Telematics, Norwegian University of Science and Technology, Trondheim, Norway, 2008.

[9] R. Åsberg. Det finns inga kvalitativa metoder och inga kvantitativa heller för den delen. *Pedagogisk Forskning i Sverige*, 6(4), pp. 270-292, 2001.

[10] N.E. Tanton. Results of a survey on television viewing distance. R&D white paper, BBC, June 2004.

[11] International Telecommunication Union (ITU). Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-11.