Frame Rate versus Spatial Quality: Which Video Characteristics Do Matter?

Korhonen, Jari; Reiter, Ulrich; Ukhanova, Ann

Published in:
Proceedings of International Conference on Visual Communications and Image Processing (VCIP'13)

Link to article, DOI:
10.1109/VCIP.2013.6706381

Publication date:
2013

FRAME RATE VERSUS SPATIAL QUALITY: WHICH VIDEO CHARACTERISTICS DO MATTER?

Jari Korhonen[1], *Ulrich Reiter*[2], *and Anna Ukhanova*[1]

[1)] Dept. of Photonics Engineering, Tech. Univ. of Denmark (DTU Fotonik)
[2)] Dept. of Electronics and Telecommunications, Norwegian Univ. of Science and Tech. (NTNU)

ABSTRACT

Several studies have shown that the relationship between perceived video quality and frame rate is dependent on the video content. In this paper, we have analyzed the content characteristics and compared them against the subjective results derived from preference decisions between different spatial and temporal quality levels. We also propose simple yet powerful metrics for characterizing spatial and temporal properties of a video sequence, and demonstrate how these metrics can be applied for evaluating the relative impact of spatial and temporal quality on the perceived overall quality.

Index Terms— Video quality, Frame rate impact

1. INTRODUCTION

Reliable assessment of video quality is essential for the development and comparison of different digital compression and processing techniques for visual information. The subjective perception of video quality is known to be dependent on several factors, of which the spatial quality (i.e. the distortion of the signal caused by compression and possibly transmission errors) is typically considered the most prevalent. Most of the objective video quality metrics known from the literature are based on solely or primarily measuring the difference between the original reference video signal and the distorted version of the same signal. This type of metrics include Peak Signal-to-Noise Ratio (PSNR), the most commonly used metric for evaluating video and image quality in the scientific community.

Less attention has been paid to the quality impact of temporal resolution. Naturally, lower frame rate implies lower perceived quality, but modeling this impact mathematically has been proven challenging. Several subjective studies have been performed to analyze the impact of different frame rates. The studies by Yadavalli et al. [1], Brun et al. [2], Huynh-Thu et al. [3] investigated the frame rate preferences in low resolution video. In spite of different methodologies and parameter sets, all these studies have shown that high spatial quality is preferred over high frame rate, especially at low bitrates. Frame rate becomes subjectively more crucial for sequences with high spatial quality and high resolution. However, the impact of frame rate on subjective perception of quality depends highly on the degree of motion in the sequences and on the content [4-6]. In general, sequences with intensive motion seem to be more sensitive to frame rate than sequences with little motion; however, the tendency is not straightforward.

The attempts to model the impact of the frame rate on the subjective quality as a function of spatial activity have shown some success [7-9], but outliers also occur, and typically the datasets used in individual studies are rather limited. A survey study of quality metric concerning temporal resolution can be found in [10]. In this paper, we present new subjective results comparing the subjective bias between spatial quality and temporal resolution. We have also analyzed how different characteristics influence this bias, in order to gain more advanced knowledge of the perception of frame rate.

2. CHARACTERIZATION OF VIDEO SEQUENCES

In our study, we have used High Definition (HD) video sequences originally provided by Technical University of Munich, downscaled to 768x432 pixels by LIVE laboratory at the University of Texas in Houston [11]. We have chosen a subset of seven sequences to represent different content and motion types.

2.1 Qualitative Characterization of Content

To understand the role of content type for the perception of frame rate and compression artifacts, we need to understand the differences in content types first. Short qualitative descriptions of each content are therefore given below.

Blue Sky: A very high contrast sequence showing dark leaves of a tree against bright blue sky, with relatively smooth rotating motion. Averagely challenging to compress, since some areas are very smooth (blue sky) and some areas are very complex (high contrast borderlines between the tree and the sky).

Pedestrian Area: A sequence with intermediate spatial complexity showing a view to a pedestrian street. There are some spots with fine details, such as pedestrians' clothing and the shops with signs, but also some smooth areas, such as the pavement of the street. The background is static (camera stands still). However, there is relatively intensive motion across the view, since there are several pedestrians and bicyclists moving in the scene.

Riverbed: A view of streaming water in a shallow river. Detailed patterns of small stones are visible below the water surface. In addition, the stream causes constantly alternating reflections of light, which makes the sequence very rich in details and fine textures, both spatially and temporally. The camera stands still, and this is why all the motion in the sequence is related to the motion of the water.

Rush Hour: A sequence showing a street in a city with heavy car traffic. There are a lot of details, but not many spots with fine detailed textures. The camera stands still and most of the motion is related to the moving cars. However, the air ripples due to heat, and this is why even the background is not perfectly static.

Station: A sequence showing a railway yard, shot from a bridge above the rails. There are areas very rich in details and textures, such as the skyline of a city in the background, but also some smoother areas, such as the sky. The camera stands in a fixed position, but it is smoothly zooming out away from the vanishing point.

Sunflower: A close-up of a bee on a sunflower. A lot of details forming a monotonic pattern, but also some smoother surfaces, especially in the upper right corner. The bee moves in different directions, seemingly randomly. The camera performs panning to different directions, apparently attempting to follow the bee.

Tractor: A sequence showing a tractor driving on a field. A lot of details and also fine textures across the whole image. Different types of motion are also present: camera pans to follow the tractor, while the tractor itself is also in an intensive motion.

2.2 Quantitative Characterization of Content

ITU has defined spatial and temporal activity indices, *SI* and *TI*, for characterizing the spatial and temporal complexity, respectively [11]. These indices, based on simple statistical properties of the analyzed video sequence, are often used for basic classification of video content types. Temporal activity metrics based on motion vectors have also been used in related work [9], but in order to avoid the computational burden of deriving the motion vectors, we have chosen to focus on ITU indices and their derivatives in our work.

The original formulation of *SI* and *TI* are given in Eqs 1 and 2:

$$SI = \max_{time}(std_{space}[Sobel(F_n)]) \quad (1)$$

$$TI = \max_{time}(std_{space}[F_n - F_{n-1}]) \quad (2)$$

In Eqs 1 and 2, F_n denotes n:th frame in the sequence, *Sobel* stands for Sobel filtering operation as defined in [12], std_{space} denotes standard deviation of the values across the spatial plane, and $\max_{time}$ denotes the maximum over the sequence of all frames.

In the literature, several modified definitions for *SI* and *TI* have been presented [7,8,13]. In those, the average value of per-frame indices is often used instead of the maximum value; this helps to avoid overemphasizing the impact of temporarily appearing objects of high motion or spatial details. In addition, standard deviation may be replaced by mean value: there is evidence that *SI* computed from mean correlates better with Kolmogorov complexity than *SI* computed from standard deviation [13]. In this paper, we have redefined *SI* and *TI* as follows:

$$SI = mean_{time}(mean_{space}[Sobel(F_n)]) \quad (3)$$

$$TI = mean_{time}(std_{space}[F_n - F_{n-1}]) \quad (4)$$

The traditional spatial and temporal activity indices are very generic measures. To characterize the video type more accurately, we propose some additional measures. We start by defining blockwise spatial and temporal activity indices $S_{n,m}$ and $T_{n,m}$ (Eqs 5 and 6), that are computed separately for each block $B_{n,m}$ of a predefined size (index n denotes the temporal position of the block, i.e. frame number, and m the spatial position index, m=1..M, where M is the number of macroblocks per frame). In this paper, we have used blocks of 16x16 pixels, similar to many image and video compression algorithms.

$$S_{n,m} = mean_{space}[Sobel(B_{n,m})] \quad (5)$$

$$T_{n,m} = std_{time}[B_{n,m} - B_{n-1,m}] \quad (6)$$

Using Eqs 5 and 6, we can define two new measures for content characterization: spatial and temporal uniformity index, *SUI* and *TUI* (Eqs 7 and 8). These values are normalized with respect to the average values of *S* and *T* in each frame.

$$SUI = mean_{time}(std[S_{n,1..M}] / mean[S_{n,1..M}]) \quad (7)$$

$$TUI = mean_{time}(std[T_{n,1..M}] / mean[T_{n,1..M}]) \quad (8)$$

Combining *SUI* and *TUI* with *SI* and *TI*, a more accurate characterization of the content can be obtained. In principle, low values of *SUI* and *TUI* indicate that spatial and temporal activity is uniformly distributed across the frames, whereas high values indicate that the activity is distributed more unevenly. Roughly, values below 0.7 can be considered as

low, and above 0.7 as high. In Fig. 1, joint interpretations for these measures are summarized.

	TUI low	*TUI* high
TI low	*Static, low motion scene*	*Mostly static, some moving objects*
TI high	*Overall motion: panning, ripple etc.*	*Overall motion with varying intensity*

	SUI low	*SUI* high
SI low	*Lot of smooth and uniform surfaces*	*Mostly smooth, some detailed objects*
SI high	*Lot of detailed patterns*	*Mix of more and less detailed patterns*

Fig. 1. Proposed content characterization.

Since the motion intensity can alternate significantly not only in the spatial plane, but also in the temporal domain, we have defined one more measure for content characterization, labeled as jerkiness index (*JI*). It is defined as standard deviation of *TI* across time, normalized with respect to *TI*:

$$JI = std_{time}(std_{space}[F_n - F_{n-1}]) / TI \qquad (9)$$

The numerical content characterizations for the test sequences are summarized in Table 1. The numerical results are corresponding reasonably well to the qualitative descriptions in Section 2.1 and the interpretations of values as shown in Fig. 1.

Table 1. Objective content characterizations

Sequence	***SI***	***SUI***	***TI***	***TUI***	***JI***
Blue Sky	80.50	0.99	31.38	0.89	0.11
Ped. Area	37.72	0.69	15.14	1.18	0.18
Riverbed	62.99	0.44	25.78	0.58	0.05
Rush Hour	33.78	0.87	8.82	1.07	0.16
Station	38.96	0.61	7.15	0.63	0.29
Sunflower	47.22	0.64	14.06	0.76	0.37
Tractor	62.70	0.51	19.31	0.50	0.08

3. SUBJECTIVE EXPERIMENT

In this study, we have used a similar methodology as described in [14], but we have applied it to higher resolution video sequences, as described in Section 2. In addition, the tests have been performed in two directions, i.e. from bad quality towards good, and vice versa. A short description of the method is given below; for more details about the methodology, the reader may refer to [7, 14].

3.1. Methodology

The methodology used for our subjective study is based on pairwise comparisons taken in steps to find the preferred path from bad quality to good quality, or vice versa. Quality is defined by two dimensions, comprising both spatial and temporal quality. At each step, the test subject can choose between two sequences, one with higher spatial quality and lower frame rate, and another with lower spatial quality and higher frame rate. After choosing the preferred sequence, the test proceeds to the next quality level, according to the user's choice. In this way, the preferred path across the spatiotemporal quality plane can be found. The methodology is illustrated in Fig. 2.

When several subjects have performed the test, it is possible to find the average path by computing the average spatial and temporal quality levels of all the subjects after each step. This is done by labeling each spatial and temporal quality level by an integer number ranging from one (lowest quality) to the number of quality levels (highest quality). Then, average position after each step can be obtained by averaging the position values in both dimensions. In [7], it has also been shown how the distribution of preferences can be turned into a quality score. Assuming that the steps between levels represent similar change in perceived quality, the average paths would follow roughly the diagonal of the spatiotemporal quality plane.

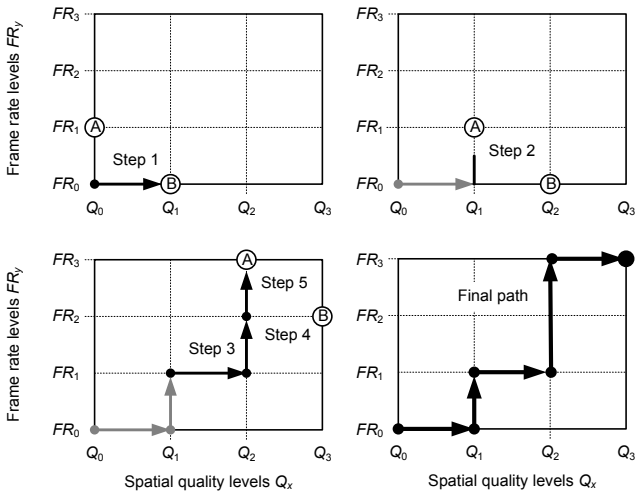


Fig. 2. Pairwise comparisons resulting in the preferred path across the spatiotemporal quality plane.

3.2. Practical study

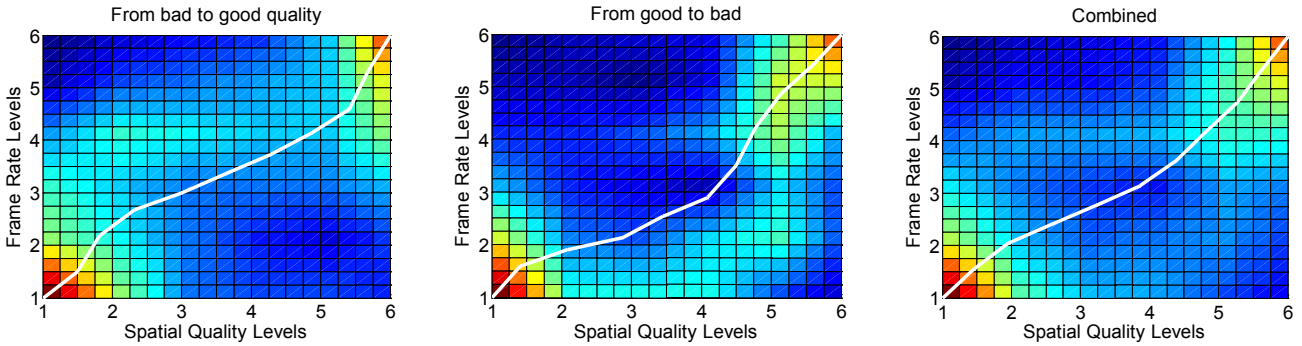
It is well known that the relationship between frame rate and perceived temporal quality is not linear but closer to logarithmic: the changes in temporal resolution are more pronounced at low frame rates than high frame rates [3]. For our experiments, we have tried to choose uniform intervals for spatial and temporal quality with respect to the observed quality differences. The frame rate levels are 24, 12, 8, 6, 4 and 2 frames per second, and different frame rates are

obtained from the full 24 fps sequences simply by skipping frames.

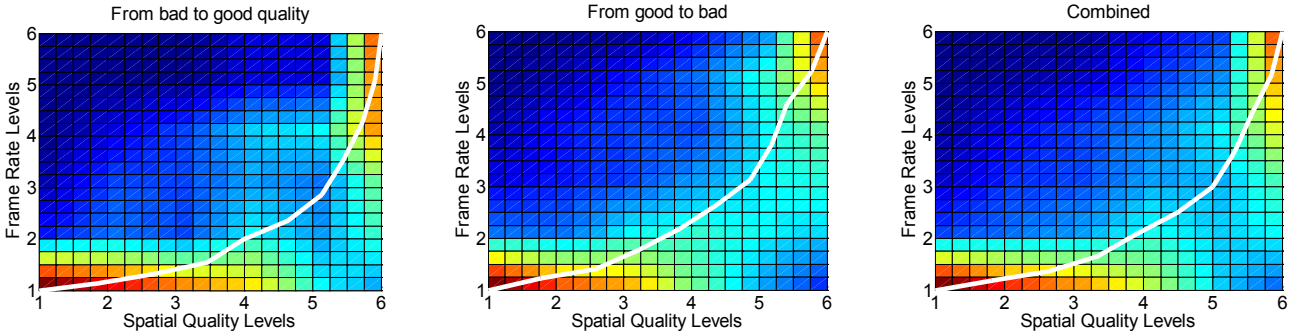
Different spatial quality levels have been generated by compressing the original sequence with H.264/AVC codec using different fixed quantization parameters (QP). To avoid quality fluctuation, we have used intra-frames only (temporal prediction has not been used). QP values have been chosen to represent the range from very bad quality (PSNR below 25 dB) to very good quality (PSNR over 40 dB). Even though we attempted to produce perceptually uniform steps, some nonlinearity may still be present; this must be taken into account, when analyzing the results.

A total of 22 subjects participated in the experiment (19 males and 3 females). The average age was 34.4 years (SD 8.6). All subjects had normal or corrected to normal vision. The average time each subject spent in the experiment, excluding training, was around 23 minutes (SD 10). An experimental session consisted of 14 trials: each of the seven different contents was evaluated by each subject from bad to good quality, and from good to bad quality. Presentation order of trials was randomized in a Latin Square design.

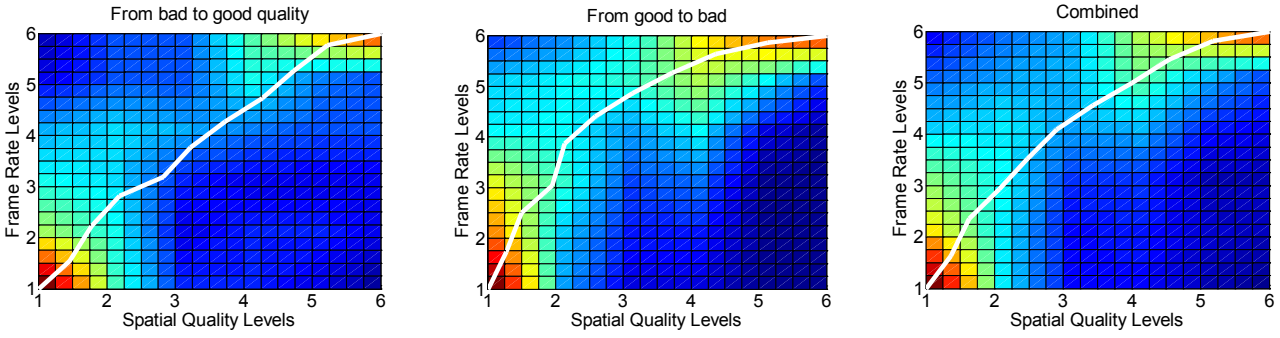
The experiment took place in a laboratory especially designed for audiovisual quality testing, featuring controlled artificial daylight conditions and a wallwash of soft light



a) *Tractor* sequence.



b) *Sunflower* sequence.



c) *Blue Sky* sequence

Fig. 3. Example paths for sequences a) *Tractor*, b) *Sunflower* and c) *Blue Sky*.

behind the screen. Content was presented on a 19" TFT with mid-grey background as recommended in ITU-R BT.500 [15]. Viewing distance was approximately 4 times the height of the screen, but could be adjusted by the participants.

4. DATA ANALYSIS

As an output of the subjective experiment, each test subject produces a path between the lowest spatial and temporal quality and the highest spatial and temporal quality. The average path can be produced by averaging the positions of each test subject after each step. We have attempted to choose the spatial quality and frame rate levels so that each step represents a uniform change in subjective quality along both axes. Therefore, if changes in temporal and spatial quality are weighted subjectively equally at each point, the average path would follow the diagonal across the spatiotemporal quality plane.

Example results are shown in Fig. 3. for *Tractor, Sunflower* and *Blue Sky* sequences (starting from bad quality towards good, good to bad, and both results combined). Levels are numbered from one to six, in increasing order of spatial quality and frame rate. The figure shows also interpolated heatmaps illustrating the density of test subjects in each area along their routes. It is worth noting that the highest density areas in the heatmaps do not match accurately with the average paths, suggesting that there are rather large differences between median paths and average paths.

In order to quantify the average preference bias between temporal and spatial quality, we have defined a simple metric, based on the average distance between diagonal and the average position after each step (endpoints excluded). Negative values denote higher emphasis on spatial quality (i.e. the average path goes below the diagonal) and positive values higher emphasis on frame rate (i.e. the path goes above the diagonal). The results are summarized in Table 2.

Table 2. Preference bias between spatial quality and frame rate. Negative values indicate that users prefer high spatial quality over high frame rate, and positive values vice versa

Sequence	bad to good	good to bad	combined
Blue sky	0.296	0.804	0.550
Ped. area	0.129	0.521	0.325
Riverbed	-0.193	0.495	0.151
Rush hour	-0.078	-0.257	-0.167
Station	-0.630	0.154	-0.238
Sunflower	-1.041	-0.778	-0.910
Tractor	-0.141	-0.341	-0.241

In most cases, test subjects seem to have stronger preference on high frame rate when they start from good quality, than when they start from bad quality. This indicates that more attention is paid to temporal resolution when the spatial quality is good, and assumedly there is a memory effect influencing each choice even after the first step. The result is in line with the related research [1,2]. *Rush Hour* and *Tractor* sequences are exceptions from this observation. For them, the paths from good to bad quality and vice versa are closer to each other than for the other sequences.

In the next phase, we attempted to predict the combined bias values as listed in Table 2, using the objective content characterization indices listed in Table 1. Several related studies have shown that high temporal activity tends to indicate higher importance of frame rate, and this is why we have mainly focused on the temporal characteristics. In Fig. 4, the quality bias values versus *TI* are plotted. The plot shows clear positive correlation between bias and *TI*, but there are also significant outliers, most notably *Pedestrian Area* and *Sunflower*.

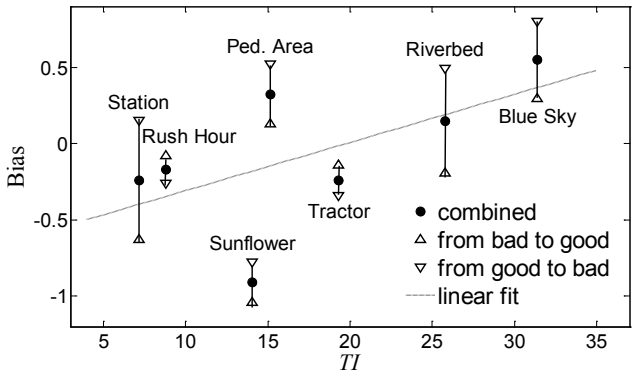


Fig. 4. Dependency between *TI* and spatial/temporal bias.

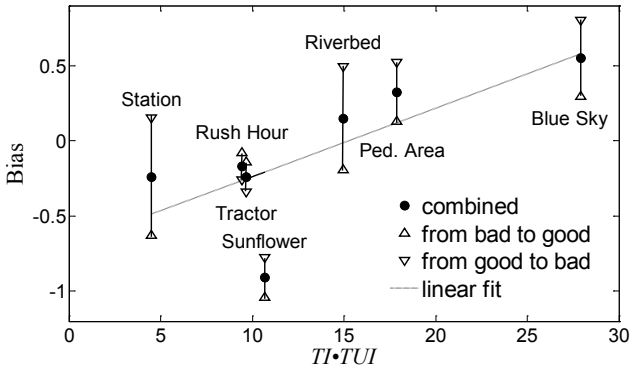


Fig. 5. Dependency between *TI·TUI* and spatial/temporal bias.

In order to improve estimation, we have tried also other ways to use the indices for prediction. It can be assumed that in case of panning or other overall motion, *TI* overemphasizes the perceived intensity of the motion. On the other hand, local intensive motion may be underemphasized, respectively. This is why combining *TI* and *TUI* by simple multiplication is an appealing alternative

to measure the overall perceived motion. Indeed, as Fig. 5. shows, the prediction accuracy can be improved by predicting bias from *TI·TUI*. Without one significant outlier, *Sunflower*, the linear fit could be improved even further.

Jerkiness of the outlier sequence, *Sunflower*, is the most significant feature that distinguishes it from the other sequences: *JI* for sunflower is 0.37, whereas the average *JI* for the other sequences is only 0.15. By intuition, it is reasonable to assume that the high emphasis on spatial quality is related to jerkiness, since natural jerkiness of motion could mask the impact of low frame rate. However, more subjective experiments should be made to confirm this hypothesis.

5. CONCLUSIONS

In this paper, we have studied the relative importance of spatial quality and frame rate on perceived quality of video sequences roughly of standard definition TV resolution. We have observed that there is a strong correlation between temporal activity level and perceived importance of frame rate. The preference bias between spatial quality and temporal resolution can be predicted reasonably accurately by using temporal activity information. The observations from our study could be used to develop more accurate metrics for evaluating both spatial and temporal quality components of video sequences. However, the number of test sequences in our study is not sufficient for the development of such a metric, and larger datasets would be required for more rigorous analysis of different influencing factors, such as jerkiness of motion.

At the same time, the proposed content characterization based on spatial and temporal activity indices on the one hand, and spatial and temporal uniformity indices on the other, allows for a more detailed quantitative characterization, revealing more and meaningful differences between videos than traditional measures alone. In the future, it might be interesting to expand this study to include the encoding case, in which temporally predicted frames are involved.

6. REFERENCES

[1] G. Yadavalli, M. Masry and S. Hemami, "Frame rate preferences in low bit rate video," in *Proc. ICIP*, pp. 441-444, Barcelona, Spain, Sep. 2003.

[2] P. Brun, G. Hauske, and T. Stockhammer, "Subjective Assessment of H.264/AVC Video for Low-Bitrate Multimedia Messaging Services," in *Proc. ICIP*, pp.1145-1148, Singapore, Oct. 2004.

[3] Q. Huynh-Thu and M. Ghanbari, "Temporal Aspect of Perceived Quality of Mobile Video Broadcasting," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 641-651, 2008.

[4] D. Wang, F. Speranza, A. Vincent, T. Martin and P. Blanchfield, "Toward Optimal Rate Control: A Study of the Impact of Spatial Resolution, Frame Rate, and Quantization on Subjective Video Quality and Bit Rate," in *Proc. VCIP*, vol. 5150, pp. 198-209, Lugano, Switzerland, Jul. 2003.

[5] J. McCarthy, M. A. Sasse, and D. Miras, "Sharp or Smooth: Comparing the Effects of Quantization vs. Frame Rate for Streamed Video," in *Proc. CHI*, pp. 535-542, Vienna, Austria, Apr. 2004.

[6] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-Dimensional Perceptual Quality Assessment for Low Bit-Rate Videos," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1316-1324, Nov. 2008.

[7] A. Ukhanova, J. Korhonen, and S. Forchhammer, "Objective Assessment of the Impact of Frame Rate on Video Quality," in *Proc. ICIP*, pp. 1513-1516, Orlando, Florida, USA, Sep. 2012.

[8] Y. Peng and E. Steinbach, "A Novel Full-reference Video Quality Metric and Its Application to Wireless Video Transmission," in *Proc. ICIP*, pp. 2517-2520, Brussels, Belgium, Sep. 2011.

[9] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual Quality Assessment of Video Considering Both Frame Rate and Quantization Artifacts," *IEEE Trans. Circuits and Syst. for Video Tech.*, vol. 21, no. 3, pp. 286-298, Mar. 2011.

[10] J.-S. Lee, F. De Simone, T. Ebrahimi, N. Ramzan, and E. Izquierdo, "Quality Assessment of Multidimensional Video Scalability," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 38-46, April 2012.

[11] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video," *IEEE Trans. Image Proc.*, vol. 19, no. 6, pp. 1427-1441, Jun. 2010.

[12] ITU-R Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," International Telecommunication Union, Geneva, Switzerland, 1999.

[13] H. Yu, and S. Winkler, "Image Complexity and Spatial Information," in *Proc. QoMEX*, pp. 12-17, Klagenfurt, Austria, Jul. 2013.

[14] J. Korhonen, U. Reiter, and J. You, "Subjective Comparison of Temporal and Quality Scalability," in *Proc. QoMEX*, pp. 161-166, Mechelen, Belgium, Sep. 2011.

[15] ITU-R Rec. BT.500-11, "Methodology for the Subjective Assessment of the Quality of Television Pictures," International Telecommunication Union, Geneva, Switzerland, 2002.