QUALITY ASSESSMENT OF MPEG-4 AVC/H.264 AND HEVC COMPRESSED VIDEO IN A TELEMEDICINE CONTEXT

A. Chaabouni[1] J. Lambert[3], Y. Gaudeau[1,2], N. Tizon[4], D. Nicholson[4] and J.-M. Moureaux[1]

[1]Université de Lorraine, CRAN, UMR 7039, 9 Avenue de la Foret de Haye Vandoeuvre les Nancy, 54500, France
e-mail: amine.chaabouni@univ-lorraine.fr,, jean-marie.moureaux@univ-lorraine.fr

[2]Université de Strasbourg, 30 Rue du Maire Andre Traband, Haguenau, 67500, France
e-mail: yann.gaudeau@unistra.fr

[3]Institut Mines Telecom, 37-39 Rue Dareau, 75014, Paris, France
e-mail: julien.lambert@imt.fr

[4]VITEC, 99 Rue Pierre Semard, 92320 Châtillon
e-mail: nicolas.tizon@vitec.com, didier.nicholson@vitec.com

ABSTRACT

To meet doctors' needs to store and share medical data remotely, lossy compression seems, today, to be an appropriate solution to manage the huge amount of these medical data. However, as there is a risk to lose critical medical information, experts' subjective quality reviews should be considered with respect to compression efficiency. In this context, we try to compare the last two video encoding international standards performances, taking into account quality assessment issues. Results show us that HEVC is more efficient than MPEG-4 AVC/H.264, offering up to 54% bit-rate saving comparing to AVC/H.264. Besides, we showed that ENT medical videos can be advantageously encoded in SD instead of Full HD resolution for low bit-rate applications. Finally, by comparison with doctors' perception, the appropriate objective metrics MSE, NQM, SSIM and MSSIM validate the previous results and confirm the superiority of HEVC over MPEG-4 AVC/H.264. They are very promising for telemedicine applications, especially in low bit-rate context.

***Keywords*— Objective and subjective quality assessment, MPEG-4 AVC/H.264/HEVC encoding standard, SD/Full HD resolution, biomedical image processing.**

1. INTRODUCTION

Nowadays, the development of telemedicine is becoming more and more important, in particular to face the problem of distance and related costs between local hospitals and reference hospitals but also for other multiple reasons as for example to maintain patients at their home. Thanks to this solution, doctors can provide remotely a high quality of care through remote consultations, tele-radiology and remote monitoring for example. In addition, telemedicine offers an efficient tool for health practitioners who want to share expertise during remote medical boards and diagnosis, especially for difficult cases. To achieve these different scenarios, it is necessary to store and transmit huge medical data like medical video streams and images over high bandwidth networks, but also, over low bandwidth ones, especially in rural areas. Then, as for example an original Full High Definition (FHD) endoscopic stream is encoded at about 2 Gbits/s, it is essential to be able to compress it to ensure a real time transmission, while maintaining a sufficient quality, for regular use by health professionals. If lossless compression was used for medical applications since it preserves the data integrity, its low performance (in terms of bit-rate) is not adapted to these applications. On the other hand, many works [1, 2] have shown that medical images (and video data) are tolerant to lossy compression under the condition that distortion due to compression is controlled. Thus, the Canadian Association of radiologists (CAR) as well as the American College of Radioloy [3], recommend the use of lossy compression in the medical context under some specific conditions.

The most appropriate way to evaluate the quality of medical compressed data with respect to their usage, consists of performing subjective tests in order to find a compromise between compression effectiveness and the experts' perception of the medical data quality after compression. A first study [4, 5] has been conducted by part of the authors relying on the European Celtic Plus project HIPERMED (High PERformance teleMEDicine platform[21]), dealing with the problem of quality assessment for AVC compressed video sequences. Here we propose to extend this study to the case of HEVC new standard and to other sequences.

In addition to the high quality telemedicine service (HIPERMED), available in hospitals provided with very good network infrastructures, the European Celtic Plus E3 (E-health services Everywhere and for Everybody) project aims at developing a web based platform to provide a wide range of telemedicine services, which will be accessible as well from consumer electronics devices.

A wide accessibility is targeted, especially in degraded network conditions. The service should remain available while keeping an acceptable QoE level. Regarding the video coding aspects, this scalability requirement involves having bitrate adaptations capabilities in order to provide the best quality for a given constrained bandwidth. In order to adapt the compressed video bitrate, different approaches can be considered. Classically, three parameters are expected to be

[1] http://hipermed.eu

[2] PROMETEE : PeRceptiOn utilisateur pour les usages du MultimÉdia dans les applications mÉdicalEs (User perception for multimedia medical usages). The platform is located in Telecom Nancy engineering school, University of Lorraine (France).

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

used for bitrate control purposes: the spatial resolutions, the frame rate and the quantization step.

In the videoconferencing context, due to the low latency constraint, the frame rate should remain as high as possible, especially in ENT endoscopic surgeries. Thus, when facing decreased transmission conditions, the adaptation mechanism has to decide either to decrease the spatial resolution or to increase the quantization factor. Hence, in order to help implementing the best strategy, this study provides comparative results of subjective evaluations obtained from medical video content compressed at different resolutions (SD versus FHD) and different quantization levels.

In a close context, rate quality performance of HEVC is assessed with respect to medical ultrasound videos [6]. In our study, we try to show HEVC video quality improvements compared to MPEG-4 AVC/H.264, for both FHD and SD resolutions. This paper is organized as follows: The different tools and methods used during this study for quality assessment are presented in Section II. Section III is dedicated to experimental results. Finally, we conclude this study in Section IV.

2. MATERIAL AND METHODS

2.1. MPEG-4 AVC/H.264 vs HEVC encoding

This study presents a comparison between the performance of the two last encoding standards HEVC and AVC/H264.

2.1.1 Brief description of both standards

In the designed HIPERMED platform, the different streams were compressed using the standard MPEG-4 AVC/ITU-T H.264 [7] (note: AVC official acronym will be used later in this document) due to its performances in video encoding comparing to the previous standards. It is currently the most used standard in network and transport applications. AVC has been developed jointly by ITU-T and ISO/IEC and it is the product of a partnership effort known as Joint Video Team (JVT). In HIPERMED platform, we use the encoder x264, which allows real time applications.
Considered as AVC successor, the new encoding standard HEVC [8] (High Efficiency Video Coding) is expected to improve the trade-off between compression ratio and visual quality. It was developed and finalized on January 2013 by the JVT team. It can support the ultrahigh resolutions 4k (3840 x 2160) and 8K (7680 x 4320) and allows parallel processing, taking benefits from multi-core architectures. This standard is based on the Coding Tree Unit (CTU) with larger macroblock sizes ranging from 16 to 64, offering more efficiency and flexibility compared to AVC that uses smaller blocks (16x16, 16x8, 8x16, 8x8, 8x4, 4x8, 4x4). In addition, it applies 3 filters: Deblocking filter, Sample offset (SAO) and Adaptive Loop Filtering (ALF), unlike its predecessor that applies only the deblocking filter. Thus, it decreases more the artifacts for the reconstructed frames. Instead of using different entropy coding, HEVC uses only the CABAC one. In our study, we use x265 encoder, the main open source implementation of HEVC which can be configured to run in real time on latest x86 architectures, released under the terms of the GNU GPL, the same license as x264 [9].

2.1.2. Encoding AVC/H.264 *and HEVC parameters*

Thanks to Dr. Gallet, an ENT (Ear, Nose and Throat) specialist, working in Nancy University Hospital France, we selected 4 original ENT sequences, acquired from a Storz endoscopic camera with S1 camera head and image control unit Image 1 HUB. The sequences are related to real ENT surgeries and identified as critical, as far as quality is concerned. These sequences last 10 seconds and are originally encoded at 1.99 Gbits/s in a full HD resolution (1920x1080 – 1080p60 – 4:2:2 – 8 bits). Two of the 4 original videos (sequence 1 and 2 of figure 1) were submitted as medical imaging reference sequences for HEVC development by the Joint Collaborative Team on Video Coding (JCT-VC) [10]. They are presented in figure 1.
To be compatible with real time/ latency constraints for both AVC and HEVC encoding, we based ourselves on the compression parameters, summarized in table 1.

Table 1. Encoding configuration

Parameter	Pixel Format	Resolution	Frequency
Value	uyvy422	1920x1080/ 720x404	60
Parameter	**Intra-refresh**	**Key-int**	**Latency**
Value	1	60	zerolatency

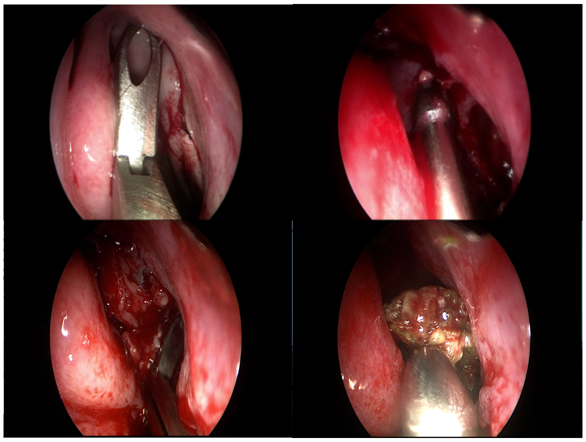

Fig. 1. Original ENT endoscopic video sequences denoted 1 to 4 clockwise from the top left (original bitrate: 1.99 Gbits/s)

2.2. Quality assessment presentation

2.2.1. Subjective tests protocol

Subjective tests performed with a panel of experts are essential to evaluate the impact of post-processing on medical videos, especially for sensitive applications such as diagnosis or surgery. Here, we propose to follow the ITU-BT.500-13 (from the International Telecommunication Union) [11] protocol, which provides methodologies for the assessment of picture and video quality, including general methods of test, the grading scales and the viewing conditions. Based on this standard, it is recommended to perform the double-stimulus continuous quality-scale (DSCQS) [11] assessment of the quality method, using a continuous scale as shown in the figure 2.

Bad
Poor
Fair
Good
Excellent

Fig. 2. Continuous scale: GUI notation application on tablet (Bad=0, Excellent=1)

2.2.2. Subjective quality assessment tools

To perform subjective tests, we used the "living lab" PROMETEE[2] an innovation platform allowing to study and well manage the technical

[2] PROMETEE : PeRceptiOn utilisateur pour les usages du MultimÉdia dans les applications mÉdicalEs (User perception for

quality of videos with respect to the medical usage. This platform, well equipped and arranged, provides a highly efficient environment to comply with the general viewing conditions for subjective assessments in such laboratory environment fixed by the ITU-BT.500-13 recommendation [11]. Thus, subjective rating sessions have been conducted in this “living lab”, where doctors were watching medical videos (encoded at 5 different AVC/HEVC compression ratios and FHD/ SD resolutions) on a standard FHD 42” screen and they give marks to each video (original and encoded sequence) through a digital tool on a tablet, during 2x24 minutes. For more information about test process, you can refer to works [4, 5].

2.2.3. Analysis of the database observers

As recommended in standard ITU-BT.500-13 [11], some sequences have been doubled and randomized. They allow us to make an initial assessment on the consistency of observers and, if the same person during the same session emits too different notations for the same sequence, it will be rejected. A second test [11] is then performed for the remaining people in order to normalize their ability to answer coherently with respect to the entire panel. If it turns out that an observer answers systematically differently than the panel, it will also be rejected.

Once these steps ended, we have a database containing the ratings of observers judged consistent. This allows us to define a MOS (Mean Opinion Score) (1), representing for each video sequence the average score of observers. MOS is given by:

$$\bar{u}_{jk} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} u_{ijk} \quad (1)$$

where N_{obs} is the number of observers and u_{ijk} is the mark of observer i corresponding to AVC/HEVC compression ratio j of the video sequence k.

This average opinion score is the unit of subjective perception of quality obtained for a panel of observers who have realized a strictly identical test. In DSCQS context, we calculated the DMOS (differential MOS), i.e. the difference between assigned to the original sequence and the note given to the encoded ones.

2.3. Objective quality assessment

Despite the relevance of subjective tests in the quality assessment of the medical videos, this method is still considered very expensive, time and human resources consuming. As an alternative, we can employ an appropriate objective metrics, which should be highly correlated to the human perception, here to the medical expert perception.

Until recently, these metrics were limited to simple ones with low performance such as PSNR. Research conducted recently has led to the establishment of psychovisual-based tools that have helped to better understand the behavior of the HVS and refine associated models. In fact, high efficient metrics for objective quality assessment have emerged in recent years like SSIM [12], PSNR-HVS [13] and HDR-VDP [14] for example. In this study, we compare a set of objective quality metrics to the participants’ scores collected during subjective tests.

Most of the above cited metrics are available in Matlab Metrix Mux library [15]. Some recent efficient metrics have been added to this library like PSNR-HVS [13] and PSNR-HVS-M [13], which attempt to model the HVS. Finally, we implemented metrics without reference (BRISQUE [16] and NIQE [17]), which allow to define the presence of compression artifacts as the traditional “blocking effect” related to the implementation of the DCT (discrete cosine transform). The reader can refer to [18] for more details on all of these metrics. In section 3, we will present results given by them.

3. EXPERIMENTAL RESULTS

Our study is based on a panel of 16 observers (5 women and 11 men) from different ENT experiences in medical curriculum (intern, extern, resident, doctor, professor). Table 2 summarizes the observers’ data in terms of years of experience.

Table 2. Features observers of our study

Years of experience	[1; 5]	[6; 10]	[11; 25]
Number of observers	7	4	5

Note that in the experiment, we founded two observers not coherent with the rest of the panel. Thus, in the following, the results rely on the 14 other observers.

3.1. x264 vs x265 quality performance

First, we focus on FHD video resolution. To measure the subjective quality of the encoded videos, we represent the curve of the evolution of the MOS score with respect to the AVC/HEVC compression bit-rate. We interpolated points (doctors’ subjective notes) using “pchip” (Piecewise Cubic Hermite Interpolating Polynomial) function as depicted in figure 4. Then, we determine the threshold of quality by choosing MOSmin= +0.1 (10% of the rating scale), a minimum value estimated as a variation that does not alter the technical quality for medical use. In other words, we consider that observers tolerate the “medical quality” of encoded video when MOS value is less than 10%. Thus, this value allows us to find the minimum AVC/HEVC compression bit-rate that can be used to encode this type of medical video.

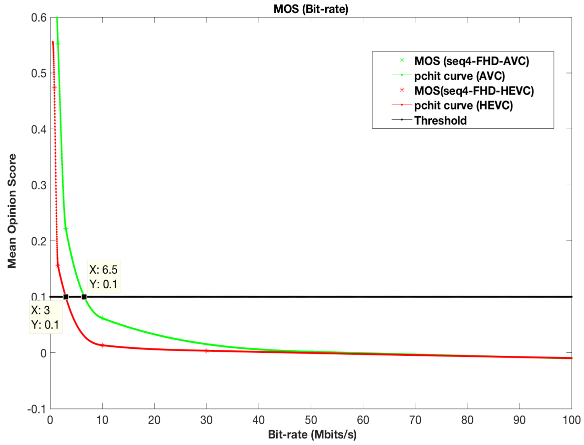

Fig. 3. MOS vs compression ratio (AVC/HEVC) (FHD sequence 4 of fig. 1)

As shown in figure 3, we note that HEVC video encoding over performs AVC in terms of quality. Fixing MOS= +0.1 as a threshold leads to 6.5 Mbits/s as AVC compression bit-rate limit. At the same MOS threshold, we can see that HEVC offers a compression bit-rate equal to 3 Mbits/s, representing about 50% of bit-rate save with respect to AVC. The different compression bit-rate thresholds for all sequences are presented in table 3.

multimedia medical usages). The platform is located in Telecom Nancy engineering school, University of Lorraine (France).

Table 3. Compression bit-rates (FHD) for MOS=+0.1

Sequence	1	2	3	4
HEVC thresholds (Mbits/s)	4.19	6.53	5	3
AVC thresholds (Mbits/s)	5.15	7.49	7.3	6.5

3.2. SD vs Full HD quality behavior

Besides the previous results, this subjective test allowed us to compare the quality behavior of HEVC / AVC compression on FHD / SD resolution videos. Due to the ITU-BT.500-13 protocol time constraints (<1 hour), the test was done only for sequence 1, judged by surgeons as the most relevant among the 4 sequences. We can see in figure 4 that this medical video sequence can be encoded using SD resolution instead of FHD in low bit-rate (<0.96Mbits/s). This result is very interesting for E3 project, showing that we can switch from FHD to SD resolution in low bit-rate transmission context. Switching from FHD to SD resolution will allow to reduce the encoding time processing. This is of high importance significant for mobile terminal and HEVC used in real time context.

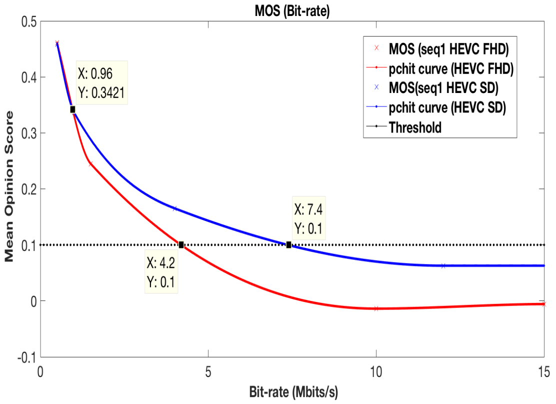

Fig. 4. MOS vs compression bit-rate (HEVC - Full HD / SD) - Sequence 1 of figure 1

3.3 Objective metrics correlation:

As we said in Section 2.3, objective metrics can be an alternative to the subjective quality assessment. However, we must find appropriate criteria to be highly correlated to the doctors' perception. Thus, we measured the correlations between objective metrics, given in Section 2.3, and MOS in order to define which of these metrics would be the most suitable for medical video quality assessment. For each of them, the Pearson correlation assessment coefficient (LCC Linear correlation coefficient) is calculated, indicating the quality of the linear regression. The closer the LCC is to 1, the better correlation between subjective and objective notes is. We put the different results of LCC in Table 4.

Table 4. Average Pearson LCC between objective and subjective measures for the 4 FHD medical video sequences

PEARSON	Avg(AVC)	Rk(AVC)	Avg(HEVC)	Rk(HEVC)
SSIM	**0,9236**	**5**	**0,9065**	**4**
UQI	0,8879	9	0,8809	6
PSNR	0,8651	10	0,8652	8
WSNR	0,8892	8	0,8839	5
VSNR	0,8215	12	0,8510	10
HDRVDP	0,9063	7	0,8720	7
IFC	0,7089	16	0,7827	16
MSE	**0,9757**	**2**	**0,9580**	**1**
MSSIM	**0,9282**	**6**	**0,9126**	**3**
NIQE	**0,9855**	**1**	0,8004	15
NQM	**0,9499**	**3**	**0,9513**	**2**
PSNRHVS	0,8363	11	0,8557	9
PSNRHVSM	0,8176	13	0,8458	11
VIF	0,7464	15	0,8136	14
VIFP	0,7951	14	0,8308	12
BRISQUE	**0,9437**	**4**	0,8263	13

Based on the table 4, we can conclude that the best objective and the most effective metrics in terms of correlation with MOS are MSE, SSIM/MSSIM and NQM in our study. In fact, MSE is a simple metric, which calculates the mean squared error between the original and the encoded video. It is more likely that doctors are more sensitive to the overall quality of the medical video (video de-noising) than to image specific structure. In fact, ENT surgeons watch the endoscopic video in real time while they operate. Thus, they seem to rather focus on the global video quality. Other specialists on the contrary, such as radiologists will need to analyze in detail every part of the image and have time to do it. NQM (Noise Quality Measure) seems to be also efficient for this kind of images, as human vision is sensitive to the variation of luminance and contrast. In addition, SSIM and MSSIM have good ranks because of their structural approach. If we compare this result to the last study [4, 5], we find the same results showing that, in addition to the last ones, NIQE (Naturalness Image Quality Evaluator) and BRISQUE are among the most correlated metrics with AVC MOS, as it measures encoding artifacts such as blocking effect. However, these two last metrics are not efficient in correlation with HEVC MOS, confirming the effectiveness of the 3 filters used by this standard.

4. CONCLUSION AND PROSPECTS

In this paper, we compared AVC/H.264 and HEVC encoding performance in terms of quality in a medical sensitive framework. As expected, the new encoding standard HEVC is more efficient than AVC/H.264. In fact, we showed that the x265 encoder (HEVC) allows reducing compression bit-rates. Indeed, in the framework of our application, for a fixed MOS=+0.1, the gain in compression bit-rate ranges between 13% and 54% with respect to x264 (AVC encoder). An additional test was performed by comparing two resolution compression behaviors, showing that we can encode medical videos using SD instead of full HD resolution in low bit-rate context (<1Mbits/s). These results are confirmed by appropriate objective metrics. As a conclusion, HEVC can be an efficient solution for low bit-rate transmission, especially for mobile networks or low bandwidth networks in rural areas. It can be thus an appropriate tool for telemedicine scenarios like the ones designed in E3 European project or more generally, for telemedicine over constrained bandwidth networks. In the next future, we intend to generalize these results to other types of medical videos.

ACKNOWLEDGEMENT

This study was conducted as a part of European Celtic E3 project. In consequence, we are deeply indebted to all E3 partners. The results of this work strongly rely on the platform PROMETEE, awarded at the Innovation program of the French Mines-Telecom Institute in 2011, which offered us the opportunity to achieve subjective tests in recommended conditions. Finally, we are very thankful to Dr. Gallet and all the employees of the ENT department of Nancy University Hospital, for allowing us to have and use endoscopic sequences from to real ENT surgeries, and to all observers who participated to this study. Namely, we would like to thank: Dr. J. Huang, Dr. G. Koch, Dr. A. Bolzer, Dr. B. Toussaint, Dr. D.T. Nguyen, Dr. T. De Saint Hilaire, Dr. J. Chauvelot, Dr. A. Bey, Dr. J. Rebois, Dr. A. Russel, Dr. C. Rumean, Dr. L. Dhaine, Dr. Y. Abu-Shama, Dr. L. Coffinet, Dr. S. Botti.

REFERENCES

[1] Nouri N, Abraham D, Moureaux J-M, Dufaut M, Hubert J, Perez M. Subjective MPEG-2 compressed video quality assessment: application to tele-surgery. In: 7th IEEE international symposium on biomedical imag- ing. 2010.

[2] Gaudeau Y, Moureaux JM. Lossy compression of volumetric medical images with 3d dead-zone lattice vector quantization. Ann Télécommun 2009;64(5–6).

[3] Canadian Association of Radiologists. Car standards for irreversible com-
pression in digital diagnostic within radiology. June 2011. p. 1–11.

[4] Chaabouni A, Gaudeau Y, Lambert J, Moureaux JM, Gallet P. Subjective and objective quality assessment for H.264 compressed medical video se- quences. In: International conference on image processing theory, tools and applications. 2014. p. 18–22.

[5] Chaabouni A, Gaudeau Y, Lambert J, Moureaux JM, Gallet P. H.264 medical video compression for telemedicine: A performance analysis. In: IRBM. 2015. p. 40-48.

[6] Razaaka M, Martinia M G, Savino K, Rate-Distortion and Rate-Quality Performance Analysis of HEVC Compression of Medical Ultrasound Videos, In: MoWNet .2014.

[7] Schwarz H, Marpe D, Wiegand T. Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans Circuits Syst Video Technol September 2007;17(9).

[8] Sullivan GJ, Ohm JR, Wiegand T. Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circuits Syst Video Technol
2012;22(12).

[9] http://x265.org & http://www.videolan.org/developers/x264.html

[10] Nicholson D, Pawałowski P, Moureaux J-M. Selected medical imaging sequences for HEVC development. In: JCTVC-O0354, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11–15th Meeting. 2013.

[11] ITU-R. Recommendation 500-13, Methodology for the subjective assess- ment of the quality of television pictures, ITU-R Rec – BT.500, 2012.

[12] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process April 2004;13(4):600–12.

[13] Wang Z, Simoncelli EP, Bovik AC. Multi-scale structural similarity for image quality assessment. In: Proceedings of the 37th IEEE Asiloma con- ference on signal, systems and computers. 2003.

[14] Mantiuk R, Kim K, Rempel AG, Heidrich W. HDR-VDP-2: a calibrated visual metrics for visibility and quality predictions in all luminance con- ditions. ACM Trans Graph 2011;30(4).

[15] Gaubatz M. Metrix Mux visual quality AssessmentPackage. Available on foulard.ece.cornell.edu.

[16] Mittal A, Moorthy AK, Bovik AA. No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 2013;20(2):209–12.

[17] Mittal A, Soundararajan R, Bovik AC. Making a "Completely Blind" im- age quality analyzer. IEEE Signal Process Lett March 2013;20(3):209–12.

[18] Moorthy A, Choi L, Bovik A, de Veciana G. Video quality assessment on mobile devices: subjective, behavioral and objective studies. IEEE J Sel
Top Signal Process October 2012;6(6).