Subjective Quality Assessment Methodology for 3D Video Compression Technology

Filip Lewandowski, Mateusz Paluszkiewicz, Tomasz Grajek, Krzysztof Wegner
Chair of Multimedia Telecommunications and Microelectronics,
Poznan University of Technology,
ul. Polanka 3, 60-965 Poznan, Poland,
e-mail: {tgrajek,kwegner}@multimedia.edu.pl

Abstract—**In this paper the methodology of subjective quality assessment for 3D video sequences is proposed. Described methodology was designed with particular attention to comparison of different 3D compression techniques. Additionally, detailed description of test session construction and design is presented. Experimental results for state-of-the-art 3D encoders performed on two 3D monitors are also included.**

I. Introduction

Currently, we observe rapid development of various kinds of 3D television services. Stereoscopic television in which a user can see 3D images is already being deployed on the market. Autostereoscopic (glassless) displays are under extensive development. Even at this moment users are using mobile devices with 3D glassless displays. First freeview television services in which user can choose scene direction viewing are currently under study.

All those 3D television services require synthesis or rendering of an intermediate views based on 3D scene representation. Currently, multiview and depth representation along with Depth Image Based Rendering (DIBR) is the most commonly used technique [1], but many different formats were proposed i.e. Layer Depth Images, Warps [2].

Development of good compression technology requires reliable quality assessment in order to balance compression performance versus provided quality. Image quality can be defined as an integrated set of factors determining the overall degree of its perfection. Obviously something else is the quality of images to be judged by suitability (e.g. medical images), and another quality understood in the context of the possible occurrence of distortion in the image. The latest is true in case of television services.

The simplest way to assess quality of an image is to show it to users and ask them about their opinion (score). This kind of quality assessment is called subjective, because it depends on the user's opinion. In order to eliminate influence of individual user deviations on the assessment, e.g. likes and dislikes, many people must be asked and their opinions (scores) should be averaged. This kind of metrics is called Mean Opinion Score (MOS). Many different procedures of subjective quality assessments have been developed over the years. They differ in:

- assessment subject (e.g. quality, distortion, fidelity of the image),
- test conditions (e.g. with or without reference),
- data processing (statistical analysis).

In order to get reliable results, all assessments should be done in precisely defined conditions. Methodology for the subjective assessment of the quality of the images has been presented in recommendation BT.500 [3]. Subjective quality assessments are the most reliable approach to judge the real quality of the image, because they base on real user experience. However, they require involvement of many people and are time consuming. In order to eliminate the need of people participation in the assessments, many automatic quality metrics have been developed. Automatic assessment is called objective (independent of individual user opinion). The simplest and the most commonly used objective quality metrics are Peak Signal to Noise Ratio of the luminance (PSNR-Y) and Mean Square Error (MSE), also calculated on luminance. Those metrics can be easily computed, processed and compared, but they can be sought only as a rough approximation of the real quality of the image. Some more sophisticated objective metrics have also been developed e.g. SSIM - Structure Similarity [4] or JND - Just Noticeable Difference [5]. However they still have many drawbacks (e.g. only a limited number of distortion types is analyzed, high computational complexity).

Because we witness the rapid development of various kinds of 3D television services, there is a strong need for reliable quality measurement procedure for 3D sequences.

Our main goal was to compare two or more compression technologies for 3D sequences. Due to the lack of good and reliable procedure of quality assessments for 3D video content, we have to develop appropriate quality assessment methodology for such a case.

II. Proposal

3D compression technology introduces different artifacts into the image, than compression of 2D images. Additionally, various representation, compression and display technologies are currently proposed and used. Therefore, the issue is what should be judged in order to get reliable comparison of 3D compression techniques. In our opinion, along with the spirit of quality assessments, quality of the views presented to the end users should be evaluated. Because many new television services utilize synthesized views, therefore, our proposition is to judge quality of the synthesized views.

Based on BT.500 recommendation we have developed original methodology for the assessment of 3D compression technology.

According to BT.500 a wide variety of basic methods may be used in television assessments. In context of 3D sequences, they can be grouped as follows:

- without a reference image,
- with a reference image:
  - which can be a rendered view obtained from uncompressed sequences,
  - which can be a view from appropriate camera (only if available).

Because we want to compare different 3D compression technologies without the influence of rendering technology used, the approach with reference view rendered from uncompressed data is more appropriate and so has been chosen.

In BT.500 5-point grading scale is recommended. We have found that in case of stereoscopic images it is insufficient to reflect the real perception of the subjects. Therefore, we propose to use 11-point scale instead, in order to better differentiate image quality. In conducted tests the sequence fidelity to the reference sequence was measured.

III. Subjects Selection

In order to credibly measure quality of the image, tests should be conducted on the broadest group of users, preferably on whole population, but this is impossible. Therefore image quality tests are conducted on limited number of subjects. Such a selected group of subjects should be a representation of the entire population. In order to carry out tests correctly, the proper selection of subjects should be ensured. Subjects should be non-experts, neither in assessing the quality of images nor in the technical aspects of digital images. Because vision system of people in age 18-30 is in optimal condition, therefore subjects should be randomly selected from this range of age. Concluding, a subject should be rather young person, who will evaluate the presented content according to personal feelings about the quality of the image.

Standard BT.500 recommends subject screening for visual acuity by Snellen charts, and proper color perception by Ishihara plates. As BT.500 is intended for evaluation of 2D images it does not include very important test for stereoscopies image perception - depth perception test. We propose the following depth perception test. A subject is shown two squares on the screen of exactly the same size (subjectively) and color at different depths. A subject is asked to point out the closer one. As squares are the same and the only difference is the depth, we can examine subject's depth perception. This test should be repeated several times with squares randomly placed in depth direction. Subjects, who have not passed at least one of the abovementioned tests must not participate in quality assessment, because their eyesight is defected and they do not represent the population average.

IV. Session Construction

In order to get statistically reliable results, whole test session/examination(-s) should be precisely designed and carried out. Test session consists of some number of test points presented one after another to the subjects. Each test point is a pair of sequences; the uncompressed reference and the processed sequence, which is the object of study.

Each test point should be presented to the subject in the following manner described in detail in recommendation BT.500.

- First, a subject is shown the number of evaluated test point (i.e. first, second, etc.) for 3 sec at mid-grey background - T1.
- Second, a subject viewes reference sequence (in our case views rendered from uncompressed data) – T2.
- Then, again 3 sec of mid-grey screen – (T3)
- Next, test point sequence (views of one of the sequences rendered from data compressed using evaluated technology) – T4.
- Finally, 5 sec of grey screen during which a subject votes (gives score for viewed test point) – T5.

This test structure is shown in Table I and it takes at least $T_P = 31$ seconds for each test point.

TABLE I
Single Test Point Design

Name	Length [s]	Type of sequence
T1	3	Grey screen
T2	at least 10	Reference
T3	3	Grey screen
T4	at least 10	Tested
T5	5	Grey screen

In order to obtain statistically reliable results (typically at 5% significance level), the appropriate number of scores for one test point have to be collected. Approximation of necessary number of scores can be evaluated based on the following equation:

$$n = \frac{t_\alpha^2 s^2}{d^2} + 1 \qquad (1)$$

where n is the needed number of scores for a single test point, t_α is quantile of Student's t-distribution, s is a standard deviation of scores and d is the confidence interval. A standard deviation of scores for a single test point can be estimated based on small preliminary viewing session. As a result, the necessary number of scores for assumed confidence level can be estimated.

At the beginning of each test session subjects learn how to evaluate the presented material, so their assessments may be unreliable. As a session is coming to an end subjects start being bored, distracted and their scores fluctuate. Scores for those test points may also be unreliable. For this reason additional k test points at the beginning and at the end of test session should be added. Scores for those test points are rejected (discarded).

Moreover, individual assessment of the subject may deviate significantly from test point to test point. In order to check how repetitive the scores given by a the single subject on the same test point are, additional l test points have to be added to each test session (these are repeated test points from the same test session). Such a test is called consistency test. The consistency test compares ratings given by each subject on the same images in the same test session. The results shall be considered consistent, if confidence intervals of the average scores of the same test point in the test session overlap. If a subject is found unreliable, their assessments must be rejected.

BT.500 recommends testing each person individually. However, in order to speed up tests and reduce their cost, it is allowed to carry them out in small groups. In order to obtain correct and reproducible results, identical conditions of observation must be ensured to each person of the test group. If this is impossible for the entire test group, it must be divided into smaller groups. Often, due to technical reasons (i.e. too small room, too few seats for subjects) test session must be repeated several times for smaller groups. If we simply repeat each test session several times we can observe contextual effect, when one of the test points in a given order can affect assessment of the next test point. In order to eliminate contextual effect each session repetition should have different presentation order of test points.

Test session should last no longer than 30 minutes. This is caused by the human eye fatigue and loss of subject's focus due to watching sequences of similar content. It can significantly affect the evaluation of sequences. Thus, if the total duration time of the test session exceeds 30 minutes (so called people focus time T_f), it has to be divided into shorter subsessions. If the presented material has to be divided into several subsessions we don't know whether the results obtained in one test subsession are comparable with all others separately. In order to check that, some number of test points from one subsession should be presented and scored in another one. Then it will be possible to check whether subjects give this repeated test point the same score. This kind of test is called session overlapping test. So, *m* additional test points have to be added to each subsession (each repeats test point from other test subsession). The session overlapping test involves comparing average results of the repeated images/test point from different test sessions. The test sessions shall be considered consistent if the confidence intervals of the average scores of the same images/test point from different test sessions overlap.

Based on the number of different test points *N* and maximum people focus time T_f we have developed equation to calculate necessary number of test session *x* which satisfies the above mentioned conditions.

$$x > \frac{N \cdot T_P}{T_f - (m + l + 2 \cdot k) \cdot T_P} \qquad (2)$$

where T_P is a single test point duration time and *m*, *k* and *l* are numbers of additional test points added as mentioned above.

At the beginning of each session, an explanation should be given to a subject about the type of assessment, the grading scale, the sequence and timing (reference picture, grey, test picture, voting period). The range and type of the impairments to be assessed should be illustrated on images other than those used in the tests, but of comparable sensitivity. It must not be implied that the worst quality seen necessarily corresponds to the lowest subjective grade. Subjects should be asked to base their judgment on the overall impression given by the image, and to express these judgments with words used to define the subjective scale [3].

V. Experiments

A. Test sequences

We have used four 3D FullHD test sequences [6-8]. These sequences are the official multiview test sequences used by an international group MPEG (Moving Picture Experts Group) which develops standards for coding audio and video. The images are also used worldwide in researches on coding, processing and quality evaluating. Table II provides a brief summary of the sequences parameters used in the subjective quality assessment tests.

TABLE II

3D Test Sequences

Name	Length	Type of sequence	Supplier
PoznanHall2	8s	natural	Poznan University of Technology
PoznanStreet	10s	natural	Poznan University of Technology
Dancer	10s	synthetic	Nokia Corporation
GTFly	10s	synthetic	Nokia Corporation

The data is taken from the Call for Proposals on 3D Video Coding Technology [9] concerning a proposed technology providing for efficient compression and reconstruction of stereoscopic images. Sequences have an average duration of 10 sec. which means on average T_P = 31 seconds for each test point to present.

B. Video encoders

In order to evaluate the proposed approach we have chosen six different 3D encoders. All of them are projects implemented on the top of state-of-the-art technique HEVC (High Efficiency Video Coding). HEVC is a draft video compression standard, a successor to H.264/MPEG-4 AVC (Advanced Video Coding)[10], currently under joint development by ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG). There were used different encoders from HEVC-3D family. Four of them are various modifications of the 3D encoder developed at Poznan University of Technology, which was proposed as a response to the Call for Proposals document on 3D Video Coding Technology [9]. This encoder was one of two best performing proposals, and it is currently subject of standardization process. In our experiments it was marked as Poznan3D Coder. Modified versions of Poznan3D Coder were described as: Poznan 3D Coder with Residual Layer Coding off, Poznan 3D Coder without Residual Layer added, MV-HEVC + Disoccluded Region Coding and HEVC + Nonlinear Depth Representation. Those modified versions utilize only some subset of available tools as described in [11] and therefore exhibit different compression efficiency. An original HEVC encoder in version 3.0 working in simulcast mode was used as a reference technique (marked as HEVC Simulcast).

C. Used monitors

In order to evaluate influence of display technology on quality assessment methodology we have used two types of 3D monitors. We have chosen best 3D monitors available on the market:

- polarization monitor: Hyundai, model S465D,
- autostereoscopic monitor: 28-view DIMENCO, model BDL5231V3D.

D. Coded material preparation

In order to evaluate coding efficiency of the investigated technology/-ies, results for wide range of bitrates need to be obtained. Three views along with three depth maps of each test sequences was encoded with all six of the encoders at some predefined bitrates. Bitrates were chosen in a way that visual quality is equally distributed from low to high. A given sequence coded with a given encoder at a given bitrate defines single test point.

TABLE III
BITRATES USED FOR TESTS

Name	Bitrate [kbps]			
PoznanHall2	140	210	320	520
PoznanStreet	280	480	800	1310
Dancer	290	430	710	1000
GTFly	230	400	730	1100

For each test point based on the decoded material we have rendered stereo pair at a spatial position located in between of spatial positions of the compressed views. Exact spatial position was selected randomly, in order to avoid optimization of the encoding technology on a given stereo pair. For autostereoscopic display we have rendered 28 dense spaced views at exact the same spatial (center of 28 views) position as randomly selected stereo pair. Those 28 views were then interleaved with software provided by display manufacturer.

As a result we obtained for each test point a video file ready to display on appropriate display.

E. Conducted tests

Prior to the tests, the necessary number of subjects was estimated using formula (1). We have conducted preliminary test session on 16 subjects in order to estimate population variance. Based on the results obtained, we have estimated the variance as $s^2 = 6.693$. The appropriate accuracy was assumed at the confidence intervals of $d = 0.55$ as s tradeoff between reliability and necessary number of subjects. For assumed significance level $\alpha = 0.05$ we estimated that at least $n = 60$ subjects are needed.

In our tests we have $N = 96$ different test points ($I = 4$ sequences, $C = 6$ encoders, $B = 4$ bitrates, as was mentioned before) with an average duration of 10 seconds.

Therefore , the total presentation time for all test points is 49min 36sec ($N{\cdot}T_P$) which is more than a human ability to focus (which is 30min as mentioned earlier).

In order to ensure test conditions and get 60 scores for each test point (see Section IV), the following steps were taken:

1. At the beginning we have calculated the number of test sessions using equation 2. The result for the given data was $x > 1.92$ so we have chosen the number of test sessions equals 2.
2. We have randomly divided all test points into two test sessions. We have 48 test points per test session. Each session lasted on average 25 minutes.
3. For each test session we have randomly selected $2{\cdot}k = 4$ test points from all test points available N and put half of them at the beginning and second half at the end of each test session.
4. We have randomly selected $l = 2$ test points from each test session and repeated them at random positions in the same test session for consistency test.
5. We have randomly selected $m = 2$ test points from each test session and added them to all others for overlapping test.
6. Because of the limitation of our test room, which can accommodate only 10 people assuring identical viewing conditions, we had to randomly divide subjects into 6 groups. Each group viewed its own version of tests sessions (randomly ordered). In other words we have redone steps 2-4 6 times. This resulted in 6 groups of tests sessions (two test sessions in each group).
7. All tests sessions were repeated separately on 2 different 3D monitors (polarization and autostereoscopic). This way a single subject has taken part in 4 test sessions.

Before each test session training of subjects was conducted. It consisted of an explanation of the session's structure, together with showing the examples of sequences with the high and low quality. The manner of assessing each sequence on specially prepared sheets was also presented. The subjects were informed about how much time they have to evaluate the image and which moment is the time for giving the score.

VI. RESULTS

After collecting all the scores from the subjects, the consistency test, session overlapping tests and screening of subjects were conducted. Both tests showed no need to reject any incorrect results.

For each test point we have calculated average score (mean opinion score) and a confidence level at assumed level of statistical significance $\alpha = 0.05$. The average level of the confidence interval was 0.33 for the results obtained for the polarization monitor and 0.34 for the results obtained for the autostereoscopic monitor. It is much better than assumed 0.55.

Fig. 1 and 2 present the Mean Opinion Score (MOS) for polarization and autostereoscopic monitors respectively.

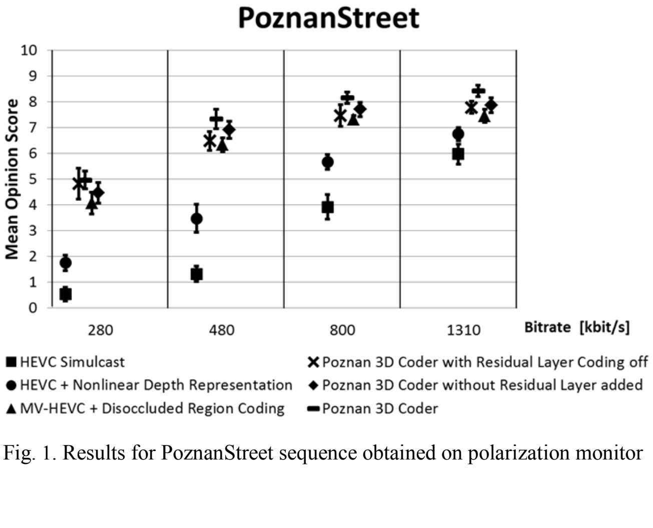

Fig. 1. Results for PoznanStreet sequence obtained on polarization monitor

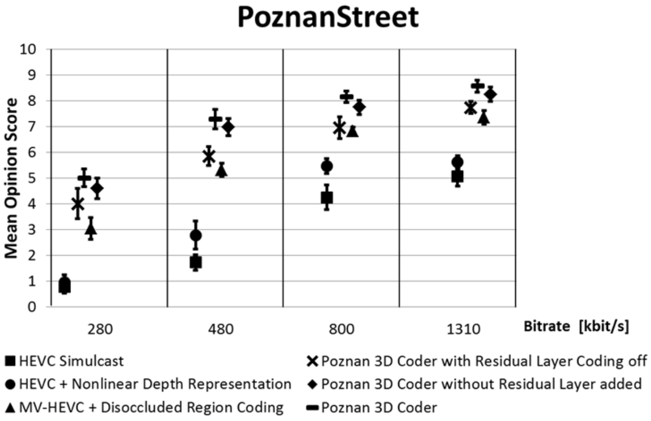

Fig. 2. Results for PoznanStreet sequence obtained on autostereoscopic monitor

Direct comparison of various coding technologies based on MOS can be tricky, because at some sequence a specific technology is better, while another technology works better with a different sequence. Therefore, we propose to use outranking charts. Outranking charts inform how many times a given encoder/technology was statistically significantly better than the others (counted over all test points). The bigger the value, the better the technology is comparing to others. Fig. 3 and 4 present exemplary outranking chart for polarization and autostereoscopic monitors respectively.

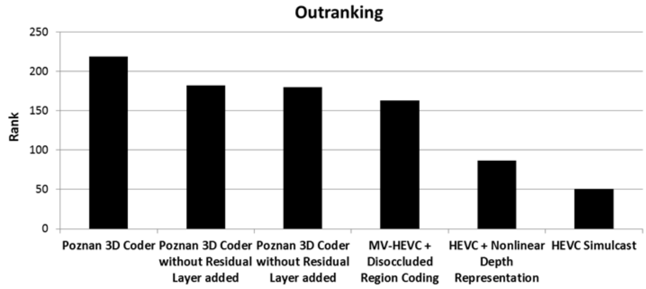

Fig. 3. Outranking chart for polarization monitor

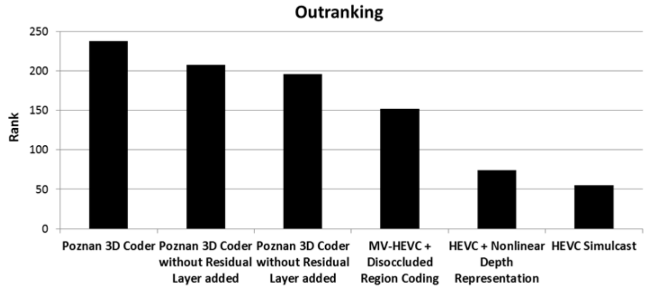

Fig. 4. Outranking chart for autostereoscopic monitor

Outranking charts show clearly that the best performing technology is Poznan 3D Coder, which has the biggest rank. The worst performing is HEVC Simulcast which has the lowest rank. The same ranking (order) of investigated technologies was obtained for both polarization and autostereoscopic monitors. It proves that the proposed methodology is independent of display technology used and can be performed on wide range of monitors

VII. Conclusions

We have proposed the methodology of subjective quality assessment for 3D video sequences based on BT 500 recommendation. Described methodology was designed with particular attention to comparison of different 3D compression techniques. Additionally, the detailed description of test session construction and design, assuring lowest time consumption with reliable results is given. Experimental results performed on state-of-the-art 3D encoders proved that the presented methodology is independent from both rendering, and displaying technology.

ACKNOWLEDGMENT

This work was supported by the public funds as a research project.

References

[1] C. Fehn, “Depth-image-based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV,” in Proc. SPIE Conf. 5291, CA, U.S.A., Jan. 2004, pp. 93–104.

[2] J. W. Shade, S. J. Gortler, L.-W. He, and R. Szeliski, “Layered depth images,” in Computer Graphics, Jul. 1998, vol. 32, Annual Conference Series, pp. 231–242.

[3] ITU-R Rec. BT.500-11, Methodology for the Subjective Assessment of the Quality of Television Pictures, 2002.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. on Image Processing, vol. 13, pp. 600–612, Apr. 2004.

[5] D.-F. Shen and L.-S. Yan, "JND Measurements and Wavelet-based Image Coding", SPIE lntemafional Optoelecmonics Expasition, July 1998.

[6] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan Multiview Video Test Sequences and Camera Parameters", MPEG 2009/M17050, Xian, China, October 2009.

[7] J. Zhang, R. Li, H. Li, D. Rusanovskyy, M. M. Hannuksela, "Ghost Town Fly 3DV Sequence for Purposes of 3DV Standardization", MPEG 2011/M20027, Geneve, Switzerland, March 2011.

[8] P. Aflaki, D. Rusanovskyy, M. M. Hannuksela, "Undo Dancer 3DV Sequence for Purposes of 3DV Standardization", MPEG 2011/M20028, Geneve, Switzerland, March 2011.

[9] “Call for Proposals on 3D Video Coding Technology,” ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, Geneva, Switzerland, March 2011.

[10] ITU-T and ISO/IEC JTC 1, "Advanced Video Coding for Generic Audiovisual Services," ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2011.

[11] M. Domański, J. Konieczny, M. Kurc, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner „3D Video Compression by Coding of Disoccluded Regions”, IEEE The International Conference on Image Processing (ICIP), Orlando, USA, 2012 – in print.