

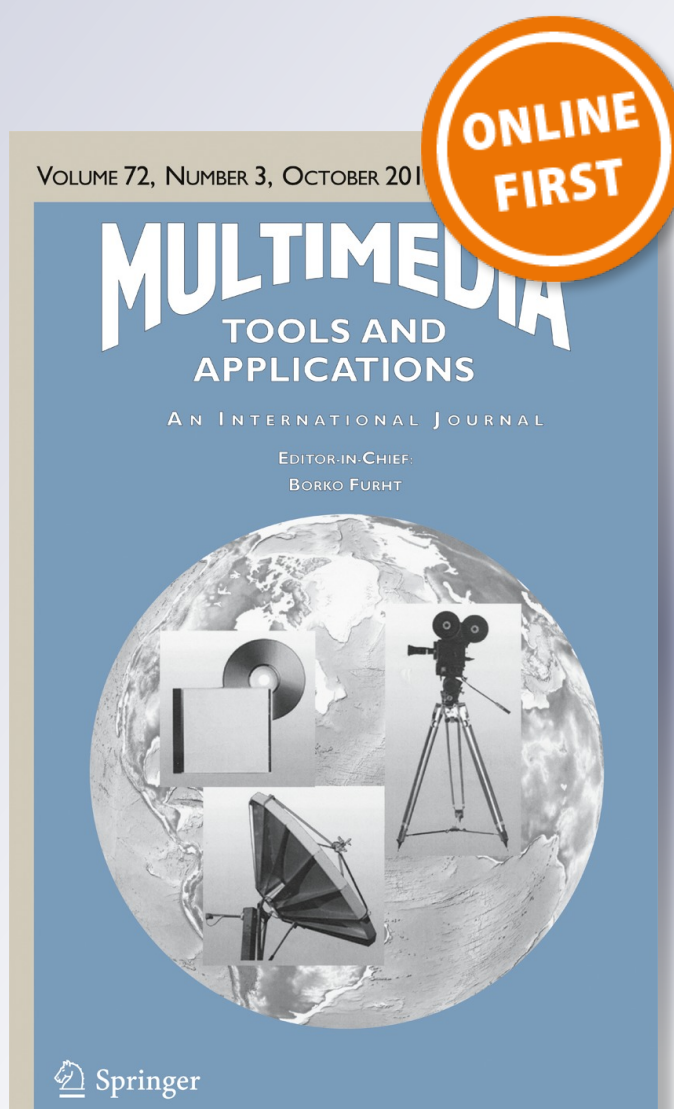
Recent developments in visual quality monitoring by key performance indicators

**Mikołaj Leszczuk, Mateusz Hanusiak,
Mylène C. Q. Farias, Emmanuel
Wyckens & George Heston**

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-014-2229-2



Your article is published under the Creative Commons Attribution license which allows users to read, copy, distribute and make derivative works, as long as the author of the original work is cited. You may self-archive this article on your own website, an institutional repository or funder's repository and make it publicly available immediately.

Recent developments in visual quality monitoring by key performance indicators

Mikołaj Leszczuk · Mateusz Hanusiak ·
Mylène C. Q. Farias · Emmanuel Wyckens ·
George Heston

Received: 31 January 2014 / Revised: 3 August 2014 / Accepted: 8 August 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract In addition to traditional Quality of Service (QoS), Quality of Experience (QoE) poses a real challenge for Internet service providers, audio-visual services, broadcasters and new Over-The-Top (OTT) services. Therefore, objective audio-visual metrics are frequently being dedicated in order to monitor, troubleshoot, investigate and set benchmarks of content applications working in real-time or off-line. The concept proposed here, Monitoring of Audio Visual Quality by Key Performance Indicators (MOAVI), is able to isolate and focus investigation, set-up algorithms, increase the monitoring period and guarantee better prediction of perceptual quality. MOAVI artefacts Key Performance Indicators (KPI) are classified into four categories, based on their origin: capturing, processing, transmission, and display. In the paper, we present experiments carried out over several steps with four experimental set-ups for concept verification. The methodology takes into the account annoyance visibility threshold. The experimental methodology is adapted from International Telecommunication Union – Telecommunication Standardization Sector (ITU-T)

The research leading to these results has received funding from the European Regional Development Fund under the Innovative Economy Operational Program, INSIGMA project No POIG 01.01.02-00-062/09-00.

M. Leszczuk (✉) · M. Hanusiak
AGH University of Science and Technology, al. Mickiewicza 30, PL-30059 Kraków, Poland
e-mail: leszczuk@agh.edu.pl

M. Hanusiak
e-mail: mateusz.hanusiak@gmail.com

M. C. Q. Farias
Department of Electrical Engineering, University of Brasília, Brasília, Brasil
e-mail: mylene@ieee.org

E. Wyckens
Orange Labs, Rennes, France
e-mail: emmanuel.wyckens@orange.com

G. Heston
Net Research, 8201 Greensboro Dr. Suite 300, McLean, VA 22102, USA
e-mail: geohes@netrsr.com

Recommendations: P.800, P.910 and P.930. We also present the results of KPI verification tests. Finally, we also describe the first implementation of MOAVI KPI in a commercial product: the NET-MOZAIC probe. Net Research, LLC, currently offers the probe as a part of NET-xTVMS Internet Protocol Television (IPTV) and Cable Television (CATV) monitoring system.

Keywords Video · Quality · Metrics · Artifacts · VQEG · Video quality · Key performance indicators · KPI

1 Introduction

In addition to traditional Quality of Service (QoS), Quality of Experience (QoE) poses a real challenge for Internet service providers, audiovisual services, broadcasters, and new Over-The-Top (OTT) services. The churn effect is linked to QoE impact and the end-user satisfaction is a real added value in this competition. QoE tools should be proactive and provide innovative solutions that are well adapted for new audiovisual technologies. Therefore, objective audiovisual metrics are frequently dedicated to monitoring, troubleshooting, investigating, and setting benchmarks of content applications working in real-time or off-line.

In the past, video quality metrics based on three common video artefacts (Blocking, Jerkiness, Blur) were sufficient to provide an efficient predictive result. The time necessary for these metrics to process the videos was long, even if a powerful machine was used. Hence, measurement periods were generally short and, as a result, measurements missed sporadic and erratic audiovisual artefacts.

In this paper, we propose a different concept known as Monitoring of Audio Visual Quality by Key Performance Indicators (MOAVI), which is able to isolate and focus investigation, set-up algorithms, increase the monitoring period, and guarantee better prediction. It is known that, depending on the technologies used in audiovisual services, the impact of QoE can change completely. So, MOAVI is able to select the best algorithms and activate or switch off features in a default audiovisual perceived list. The scores are separated for each algorithm and preselected before the testing phase. Then, each artefact Key Performance Indicator (KPI) can be analysed by working on the spatially and/or temporally perceived axes.

The classic quality metric approach cannot provide pertinent predictive scores with a quantitative description of specific (new) audiovisual artefacts, such as stripe error or exposure distortions. MOAVI is an interesting approach because it can also detect the artefacts present in videos, as well as predict the quality as described by consumers. In realistic situations, when video quality decreases in audiovisual services, customers can call a helpline to describe the annoyance and visibility of the defects or degradations in order to describe the outage. In general, they are not required to provide a Mean Opinion Score (MOS). As such, the concept used in MOAVI is completely in phase with user experience. There are many possible reasons for video disturbance, and they can arise at any point along the video chain transmission (filming stage to end-user stage).

In this paper, we present our experiments carried out over several steps with four experimental set-ups for concept verification. The methodology takes into account the annoyance visibility threshold. The experimental methodology is adapted from International Telecommunication Union Telecommunication Standardization Sector (ITU-T) Recommendations: P.800 [5], P.910 [7] and P.930 [6]. For each metric the test consisted of two parts:

setting the threshold of distortion visibility and performing the KPI checking process. Prior to the test, results of subjective experiments were randomly split into two independent sets for each part of the test. These two sets were the training set and the verification set. We also present the results of KPI verification tests. Finally, we describe the first implementation of MOAVI KPI in a commercial product, the NET-MOZAIC probe. The probe is offered by Net Research, LLC as part of NET-xTVMS IPTV and CATV monitoring systems.

The remainder of this paper is structured as follows: Section 2 is devoted to the state-of-the-art background. Section 3 discusses origins of the artefacts. Section 4 presents video artefacts and related KPI for automated quality checking. Section 5 overviews experimental set-ups for concept verification. Section 6 analyses results on KPI. Section 7 presents deployment. Finally, Section 8 summarizes the paper.

2 State-of-the-art background

This section presents limitations of current full-reference (FR), reduced-reference (RR), and no-reference (NR) metrics for standardized models. Most of the models in ITU-T recommendations were validated on video databases that used one of the following hypotheses: frame freezes lasting up to 2 seconds; no degradation at the beginning or at the end of the video sequence; no skipped frames; clean video reference (no spatial or temporal distortions); minimum delay supported between video reference and video (sometimes with constant delay); and up or down-scaling operations not always taken into account [17].

As mentioned earlier, most quality models are based on measuring common artefacts, such as blur, blocking, and jerkiness, for producing a prediction of the MOS. Consequently, the majority of the algorithms generating a predicted MOS show a mix of blur, blocking, and jerkiness metrics. The weighting between each KPI could be a simple mathematical function. If one of the KPIs is not correct, the global predictive score is completely wrong. Other KPIs mentioned in MOAVI are usually not taken into account (exposure time distortion, noise, block loss, freezing, slicing, etc.) in predicting MOS [17].

ITU-T has been working on MOAVI-like distortions for many years [6]. However, only for FR and RR approaches. The history of the ITU-T recommendations for video quality metrics is shown in Table 1. Table 2 shows a synthesis of the set of metrics that are based on video signals [17]. As can be noticed from both tables, there is a lack of developments for the NR approach.

Table 1 The history regarding ITU-T Recommendations (based on: [17])

Model Type	Format	Rec.	Year
FR	SD	J.144 [8]	2004
FR	QCIF–VGA	J.247 [10]	2008
RR	QCIF–VGA	J.246 [9]	2008
FR	SD	J.144 [8]	2004
RR	SD	J.249 [12]	2010
FR	HD	J.341 [13]	2011
RR	HD	J.342 [14]	2011
Bitstream	VGA–HD	In progress	Exp. 2013/2014
Hybrid	VGA–HD	In progress	Exp. 2013/2014

Table 2 Synthesis of FR, RR and NR MOS models (based on: [17, 24])

		Type of ITU-T Model		
		<i>FR</i>	<i>RR</i>	<i>NR</i>
<i>Resolution</i>	<i>HDTV</i>	J.341 [13]	n/a	n/a
	<i>SDTV</i>	J.144 [8]	n/a	n/a
	<i>VGA</i>	J.247 [10]	J.246 [9]	n/a
	<i>CIF</i>	J.247 [10]	J.246 [9]	n/a
	<i>QCIF</i>	J.247 [10]	J.246 [9]	n/a

In a related research, Gustafsson et al. [4] addressed the problem of measuring multimedia quality in mobile networks with an objective parametric model [17]. Closely related work are ongoing standardization activities at ITU-T SG12 on models for multimedia and Internet Protocol Television (IPTV) based on bit-stream information. SG12 is currently working on models for IPTV. Q.14/12 is responsible for these projects, provisionally known as P.NAMS (non-intrusive parametric model for assessment of performance of multimedia streaming) and P.NBAMS (non-intrusive bit-stream model for assessment of performance of multimedia streaming) [17]. P.NAMS uses packet-header information (e.g., from IP through MPEG2-TS), while P.NBAMS also uses payload information (i.e., coded bit-stream) [23]. However, this work focuses on the overall quality (in MOS units), while MOAVI is focuses on KPIs [17].

Most of the recommended models are based on global quality evaluation of video sequences, as in the P.NAMS and P.NBAMS projects. The predictive score is correlated to subjective scores obtained with global evaluation methodologies (SAMVIQ, DSCQS, ACR, etc.). Generally, the duration of video sequences is limited to 10 s or 15 s in order to avoid a forgiveness effect (the observer is un-able to score the video properly after 30 s and may give more weight to artefacts occurring at the end of the sequence). When one model is deployed for monitoring video services, the global scores are provided for fixed temporal windows and without any acknowledgement of the previous scores [17].

3 Origins of artifacts

MOAVI artifact KPIs are classified into four directories based on their origins: capturing (Section 3.1), processing (Section 3.2), transmission (Section 3.3), and display (Section 3.4).

3.1 Capturing

Capturing artefacts are introduced during video recording. Images and video are captured using cameras that comprise of an optical system and a sensor with processing circuitry. Since they occur at the front end of image acquisition, capturing artefacts are common in both analogue and digital systems. Reflected light from the object or scene can form an image on the sensor [18]. Examples of capturing artefacts include blur, exposure time distortions, and interlacing.

3.2 Processing

Processing is required to meet constraints, such as bandwidth limitations, imposed by the medium and to provide protection against medium noise. There are many coding techniques for removing redundancies in images and video and, consequently, compress them. Coding can introduce several types of artefacts, such as reduced spatial and temporal resolution, blocking, flickering, and blurring. Coding artefacts are the most common, dominant, and undesirable visible artefacts [18].

3.3 Transmission

When data is transmitted through a medium, some of the data may be lost, distorted, or may appear as repeated due to reflections. When data arrives through many paths (in addition to the direct path), the distortion is known as multi-path distortion, which affects both analogue and digital communications [18]. Example of transmission artefacts include block loss, blackout, freezing, and slicing.

3.4 Display

As the technology developed, different display systems were made available, offering different subjective qualities for the same resolution. But, with currently available display screens, the difference in quality is reduced to a minimum for the different technologies OLED, LCD, and SED. The most visible display artefacts include blackout and slicing.

4 Video artefacts and related key performance indicators for automated quality checking

This section introduces video artefacts and related KPI for automated quality checking. The described quality degradation types are: blur, exposure time distortions, noise, block loss, blocking, freezing, and slicing.

We acknowledge that the disadvantage of NR metrics is the lack of independence of visual artefacts (for instance slicing and block loss are dependent; similarly, for exposure and noise metrics).

4.1 Blur

Blur (blurring) shows as a reduction of edge sharpness and spatial detail. A sample video frame containing blur is shown in Fig. 1a. In compressed video, it results from a loss of high frequency information during coding. Blur may also appear as a result of camera movement during video recording, what is known as motion blur.

There are several metrics for measuring blur. In this work, the blur metric is based on taking the cosine of the angle between plane perpendiculars in adjacent pixels, which is known to be a good characteristic of picture smoothness [21].

4.2 Exposure time

Exposure time distortions are visible as an imbalance in brightness, i.e. the presence of frames that are too dark or too bright. A sample video frame exposure time distortion is



Fig. 1 Sample video frames containing artefacts: **a** Blur, **b** Exposure time distortions, **c** Noise, **d** Block loss, **e** Blocking, and **f** Slicing

shown in Fig. 1b. They are caused by incorrect exposure time or by recording video without a lighting device. It can be also a result of the application of digital video processing algorithms.

Mean brightness of the darkest and brightest parts of the image is calculated in order to detect exposure time distortions. The result of the calculations have range between 0 and 255. There are two thresholds which define whether an over- or under-exposure alert is to be set [20].

4.3 Noise

Physically, noise is defined as an uncontrolled or unpredicted pattern of intensity fluctuations that is unwanted and does not contribute to the good quality of a video or image. A sample video frame containing noise is shown in Fig. 1c. There are many types of noise artefacts present in compressed digital video, two of the most common are mosquito noise and quantization noise. Mosquito noise is perceived as a form of edge busyness characterized by moving artefacts or blotchy noise superimposed over objects. Quantization noise is introduced by the quantization of the pixel values. It looks like a random distribution of very small patterns (snow). It can be gray or coloured noise, but its distribution is not uniform over the image.

The noise metric estimates the noise level by calculating the local variance of the most flat (low activity) areas of the frame or image. A threshold indicates the value above which the noise artefact is noticeable.

4.4 Block loss

Block loss occurs when some of the data packets forming the compressed stream of video are lost during one of the transmission stages. As a result of this loss, one or more flat colour blocks are included in the frame in the position of the lost ones. If an error concealment algorithm is used, the lost blocks are substituted by an approximation (prediction) of the original blocks. A sample video frame containing block loss artefacts is shown in Fig. 1d.

The block loss metric relies on finding every horizontal and vertical edge in the image. Based on these edges, each macro-block is classified as lost or not. The number of lost macro-blocks indicates the visibility of the block loss distortion.

4.5 Blocking

Blocking (block distortion) is the most visible image and video degradation of all artefacts. A sample video frame containing blocking is shown in Fig. 1e. The effect is caused by all block-based coding techniques. It is a well-known fact that all compression techniques divide the image into small blocks and then compress them separately. Due to the coarse quantization, the correlation among blocks is lost and horizontal and vertical borders appear [15].

The metric for blocking is calculated only for pixels at boundaries of 8×8 blocks [21]. Its value depends on two factors: magnitude of color difference at the block's boundary and picture contrast near boundaries. The threshold determines the score below which the blocking distortion is noticeable [1].

In coders, the blocking is not limited to 8×8 . For example, even if the MPEG-2 coder works in 8×8 , the block size can be 12×8 , 10×8 and 11×8 , 16×8 , depending on the pre-filtered condition (1/2, 2/3, 3/4 or 1/1). For H.264 and HEVC coders, the blocking size can vary from 4×4 up to hundreds of pixels.

Blocking measurement at 8×8 blocks may cause problems when a flexible macro-block definition that accepts 4×4 blocks is used. We are aware of this problem. It should be noted that, in the case of NR metrics, the video encoding parameters are not known. For this reason, constant parameters (default values for typical codecs) are accepted, which is one drawback of NR metrics.

4.6 Freezing

Freezing (jerkiness, stilted and jerky motion) is frequently found in IPTV sequences during high motion scenes. It is perceived as time-discrete “snapshots” of the original continuous scene strung together as a disjointed sequence. Motion jerkiness is associated with temporal aliasing, which in turn is a manifestation of the use of an inadequate sampling/display rate than what is necessary to accommodate the temporal bandwidth of the source. The appearance of jerkiness in a sequence can also be caused by transmission delays of the coded bit-stream to the decoder. The severity of the degradation is also subject to the decoder's ability to buffer against these intermittent fluctuations.

Detection of freezing is based on a calculation of the differences between consecutive frames, which gives a measure of which temporal information activity metric is to be used.

A result, a value below the freezing threshold indicates a frozen frame. The visibility of a freezing artefact also depends on its duration. The shortest time for which viewers can perceive the presence of this artefact equals 100 milliseconds, which is estimated based on the work of van Kester et al. [16].

4.7 Slicing

The slicing artefact appears when a limited number of video lines (stripes) is severely damaged. A sample video frame containing slicing is shown in Fig. 1f. It is caused by a loss of video data packets. A decoder can reconstruct (predict) lost slices in a image based on adjacent slices.

Detection of the slicing artefact is performed by analysing image stripe similarity, both in vertical and horizontal directions. The threshold refers to the minimum value of the similarity value above which the distortion is noticeable.

5 Experimental set-ups for concept verification

This section introduces four experimental set-ups for concept verification: CONTENT 2009 experimental set-up (described in Section 5.1), VQEG HDTV 2010 experimental set-up (described in Section 5.2), INDECT 2011 experimental set-up (described in Section 5.3), and VARIUM 2013 experimental set-up (described in Section 5.4). Table 3 shows a mapping between KPI and experimental set-ups used for KPI concept verification.

The CONTENT 2009, VQEG HDTV 2010, and INDECT 2011 subjective experiments were based on MOS. During the data analysis the scores given to the processed versions were subtracted from the scores given to the corresponding reference to obtain degradation/differential MOS (DMOS). Then, the artefact perceptibility threshold was set to $DMOS = 4$, based on degradation category rating (DCR) procedure test instructions (see ITU-T P.800 [5])¹:

- 5 Degradation is imperceptible.
- 4 Degradation is perceptible, but not annoying.
- 3 Degradation is slightly annoying.
- 2 Degradation is annoying.
- 1 Degradation is very annoying.

The selection of an adequate threshold requires further investigation, since it is based on DCR and the tests were conducted using absolute category rating (ACR). Not all effects observed in ACR tests are observed with the DCR procedure and vice versa.

It must also be acknowledged that the authors are aware that the proposed methodology based on optimal thresholds detection has the drawback of ‘optimal thresholds generalization’, since the methodology is affected by the particular video material, experimental set-up, participants, and measurement session. Nevertheless, care has been taken to minimize these effects.

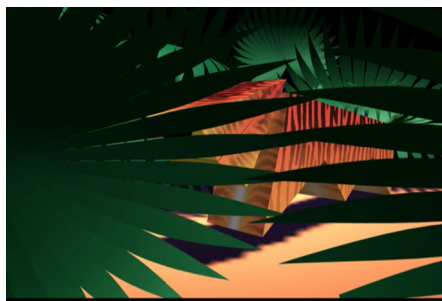
¹The original DCR instructions are related to audio testing, while MOAVI relates to visual testing. Consequently, the “audibility” term has been extended to “perceptibility”.

Table 3 Mapping between KPI and experimental set-ups

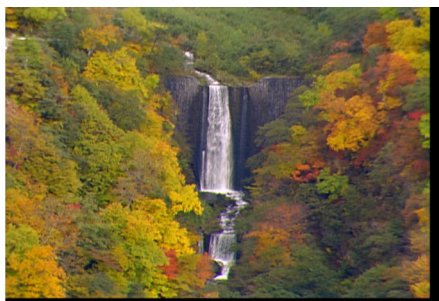
KPI	Experimental Set-up
Blur	VARIUM 2013 (Section 5.4)
Exposure Time	INDECT 2011 (Section 5.3)
Noise	CONTENT 2009 (Section 5.1)
Block Loss	VQEG HDTV 2010 (Section 5.2)
Blocking	VARIUM 2013 (Section 5.4)
Freezing	VQEG HDTV 2010 (Section 5.2)
Slicing	VQEG HDTV 2010 (Section 5.2)

5.1 CONTENT 2009 experimental setup

In the CONTENT 2009 experimental set-up, the subjective experiment was run at AGH University with 100 subjects rating video sequences affected by noise [1]. Four diverse test sequences, provided by the Video Quality Experts Group (VQEG), were used: “Betes”, “Autumn”, “Football”, and “Susie”. The video sequences reflect two different content characteristics (i.e. motion and complexity), as shown in Fig. 2. The frame-rate was 30, with a resolution of 720×486 . The subjective tests methodology was the ITU’s ACR with hidden reference (ACR-HR) described in ITU-T P.910 [7]. Following the VQEG guidelines for ACR methods, the MOS numerical quality scale (defined in ITU-T P.800 [5]) was used.



(a)



(b)



(c)



(d)

Fig. 2 Sample frames of 720x486/30fps original videos used in CONTENT [1] and INDECT [20] projects: (a) Betes, (b) Autumn, (c) Football, and (d) Susie

More information on the results from this experimental set-up has been provided by Cerqueira, Janowski, Leszczuk et al. in [1].

5.2 VQEG HDTV 2010 experimental setup

In 2010, VQEG presented the results from the HDTV validation test of a set of objective video quality models. Six subjective experiments provided data against which model validation was performed. The experiments were divided into the four 1080p and 1080i video formats. The 720p content was inserted into the experiments as test conditions, for example by converting 1080i 59.94 fields-per-second video to 720p 59.94 frames-per-second, compressing the video, and then converting back to 1080i. The original videos represented a diverse range of content, as shown in Fig. 3. The subjective experiments included processed video sequences with a wide range of quality. The impairments included in the experiments were limited to MPEG-2 and H.264, both ‘coding only’ and ‘coding plus transmission’ errors [22].

A total of 12 independent testing laboratories performed subjective experimental testing (AGH University, Psytechnics, NTIA/ITS, Ghent University – IBBT, Verizon, Intel, FUB, CRC, Acreo, Ericsson, IRCCyN, and Deutsche Telekom AG Laboratories). To anchor video experiments to each other and assist in comparisons between gathered subjective data, a common set of carefully chosen video sequences was inserted into all experiments. These common sequences were used to map the six experiments onto a single scale.

More information on the results from this experimental set-up has been provided in a technical report published in 2010 [22]. Mikołaj Leszczuk is one of the co-authors of this report, providing important technical contributions to the document.

5.3 INDECT 2011 experimental setup

The purpose of this AGH subjective experiment was to calculate a mapping function between exposure metric values and quality perceived by users [20]. Testers rated video sequences affected by different degradation levels, obtained using the proposed exposure generation model. Similarly to the CONTENT 2009 experimental set-up (described in Section 5.1), four test sequences, provided by the VQEG [19], were used: “Betes”, “Autumn”, “Football”, and “Susie”. The video sequences reflected two different content characteristics (i.e. motion and complexity). For both over- and under-exposure, six degradation levels were introduced into each test sequence. The exposure range was adjusted to cover the whole quality scale (from very slight to very annoying distortions). The applied subjective test methodology used was the ACR-HR.

More information on analysis of results from this experimental set-up has been provided by Romaniak, Janowski, Leszczuk and Papir, in [20].

5.4 VARIUM 2013 experimental setup

We considered data from an experiment of the VARIUM project, developed jointly by the University of Brasília (Brazil) and the Delft University (The Netherlands). The data and the experiment methodology are described by Farias et al. in [2]. In this experiment, subjects judged the quality of videos with combinations of blocking and blur artefacts. Seven high definition (1280×720 , 50 fps) original videos were used. The videos were all eight seconds long and represented a diverse range of content, as shown in Fig. 4. To generate sequences with blocking and blur, the authors used a previously developed system for generating

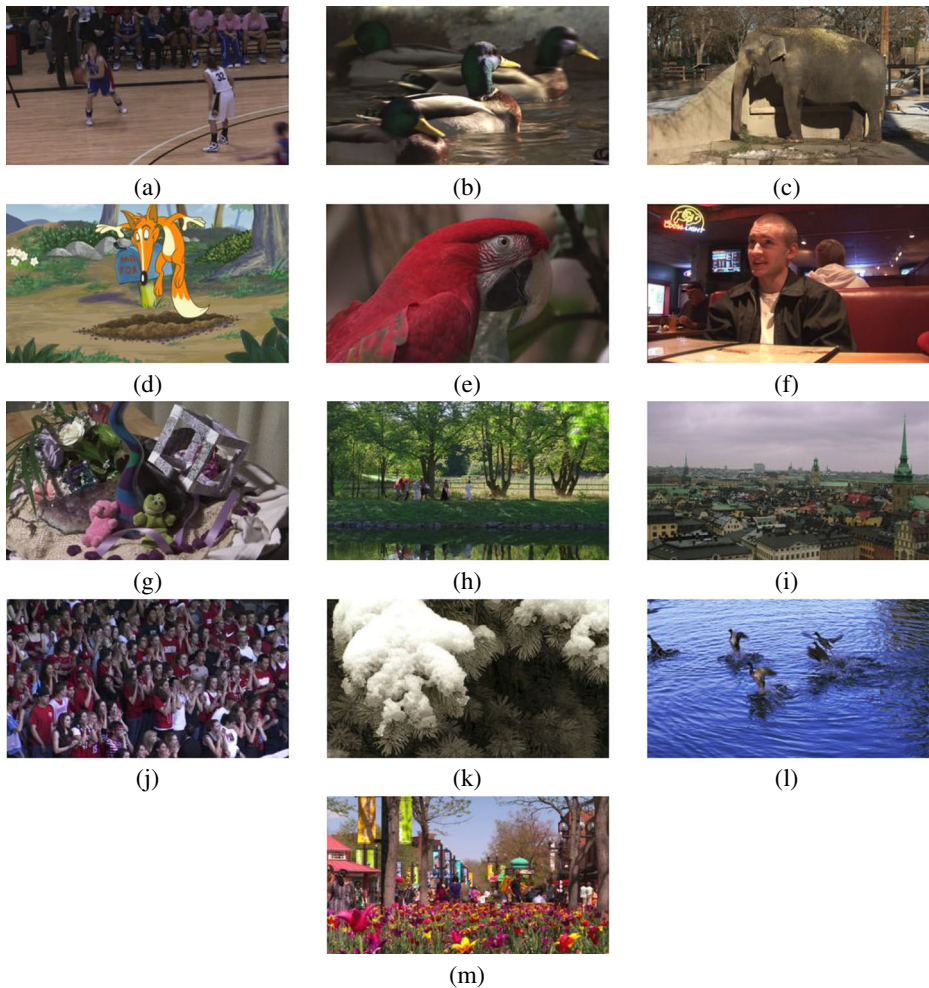


Fig. 3 Sample frames of 1920x1080/25fps original videos used in the VQEG HDTV project [22]: (a) VQEGHD 4, SRC 1, (b) VQEGHD 4, SRC 2, (c) VQEGHD 4, SRC 3, (d) VQEGHD 4, SRC 4, (e) VQEGHD 4, SRC 5, (f) VQEGHD 4, SRC 6, (g) VQEGHD 4, SRC 7, (h) VQEGHD 4, SRC 8, (i) VQEGHD 4, SRC 9, (j) Common Set SRC 11, (k) Common Set SRC 12, (l) Common Set SRC 13, and (m) Common Set SRC 14

synthetic artefacts [3], which is based on the algorithms described in the Recommendation P.930 [6]. The method makes it possible to combine as many or as few artefacts as required at several strengths. This way, one can control both the appearance and the strength of the artefacts in order to measure the psychophysical characteristics of each type of artefact signal separately or in combination.

Experiments were run with one subject at a time. Subjects were seated straight directly in front of the monitor, centred at or slightly below eye height for most subjects. A distance of 3 video monitor screen heights between the subject's eyes and the monitor was maintained [11]. All subjects watched and judged the same test sequences. After each test sequence was played, the question "Did you see a defect or an impairment?" appeared on the computer



Fig. 4 Sample frames of 1280x720/50fps original videos used in the VARIUM project [2]: (a) Joy Park, (b) Into Trees, (c) Crowd Run, (d) Romeo & Juliet, (e) Cactus, (f) Basketball, and (g) Barbecue

monitor. The subject chose a ‘yes’ or ‘no’ answer. Next, they were asked to give a numerical judgement of how annoying (bad) the detected impairment was. The subjects were asked to enter annoyance values from ‘0’ (not annoying) to ‘100’ (most annoying artefacts in the set).

Standard methods are used to analyse the annoyance judgements provided by the test subjects [11]. These methods were designed to ensure that a single annoyance scale was applied to all artefact signal combinations. More information on analysis of results from this experimental set-up is provided in [20].

6 Results of KPI

This section presents the results of KPI verification tests. For each metric the test consists of two parts: setting the threshold of distortion visibility (Section 6.1), and the KPI checking process (Section 6.2). Prior to the test, results of subjective experiments (Section 5) were randomly split into two independent sets, which correspond to each part of the test. These two sets were the training set and the verification set.

6.1 Setting the metric threshold values

For each metric, the procedure of determining the visibility threshold included the following steps:

1. For all video sequences, the metric values are calculated from the appropriate subjective experiment.

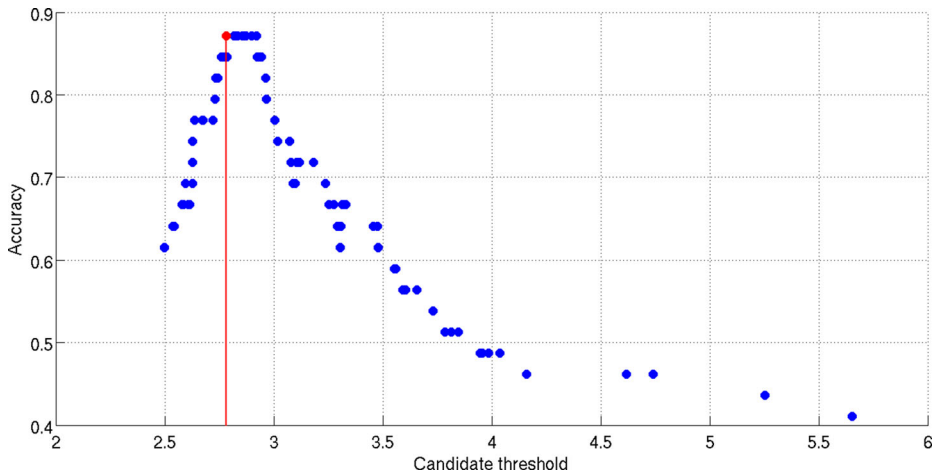


Fig. 5 Blur metric threshold determination. The points represent the relationship between candidate thresholds and accuracy. The line is drawn at the best candidate, which is chosen to be the metric threshold

2. We assume each metric value as a threshold candidate (th_{TEMP}). For values less than th_{TEMP} we set the KPI to '0'. Otherwise (values that are the same or above), we set KPI to '1'.
3. For each th_{TEMP} , we calculate the accuracy rate of the resulting assignments. The accuracy rate is given by the fraction of KPI that matches the indications given by subjects in the training set.

$$\text{accuracy}(th_{TEMP}) = \frac{\text{number of matching results}}{\text{number of results}} \quad (1)$$

4. We set the threshold of the metric as the value of th_{TEMP} with the best (maximum) accuracy. In the case of several th_{TEMP} values having the same accuracy, we select the lowest value.

Figure 5 illustrates the procedure of determining the threshold for the blur KPI. The threshold values are shown in Table 4.

6.2 KPI verification

In the second part of the test, the correctness of the KPI was checked. The accuracy of the KPI was calculated using (1), and compared against the indications from the verification set. Table 4 presents the verification results.

Thresholds themselves are strongly dependent on the operation of specific metrics, and their significance does not matter for the person not working with data metrics – they are simply the values for Key features of the NET-200 are as follows which the final (binary) metrics results are most consistent with the results of our subjective tests.

7 Deployment

This section describes the first implementation of MOAVI KPI in a commercial product: the NET-MOZAIC probe. The probe is currently offered by Net Research, LLC as

Table 4 KPI verification: probability of distortion detection and corresponding used threshold value

Metric	Probability of distortion detection	Value of threshold
Blur	0.86	2.78
Exposure Time	0.81	78 and 178
Noise	0.85	3.70
Block loss	0.84	5.3
Blocking	0.94	0.85
Freezing	0.80	0
Slicing	0.85	7

a part of NET-xTVMS [1] IPTV and CATV monitoring systems. The Net-xTVMS system offers several types of probes located at the strategic sites of the IPTV networks. One of these sites is the head-end where sources of programming content are encrypted and inserted into the servers for distribution via primary and secondary IP rings to core and edge routers, and to edge QAMs. Therefore the head-end is the only location where content is not encrypted and it can be decoded and analysed for image distortions. The only other place where that analysis could be done is the customer Set To Box output (typically HDMI). Due to large number of these Set Top Boxes the cost of such analysis would be prohibitive. In addition content quality analysis at the source prevent distribution of distorted programming.

There are many sources of image quality distortions: cameras, recording equipment, video servers, encoders, decoders, transmission disturbances etc. IPTV operators have a choice of monitoring content quality at the head-end or ignoring it. If high programming quality standards are required and the NET-xTVMS system is already present, the NET-MOZAIC is an obvious choice. However, the NET-MOZAIC can function as an independent probe even if the NET-xTVMS is not present. Therefore, for quality minded network operators NET-MOZAIC type product is always a good choice.

7.1 NET-MOZAIC probe description

The NET-MOZAIC is a PC-based IPTV test instrument and multi-channel display generator. Its primary function is decoding groups of IPTV channels (streams), performing an image quality analysis on these channels, and displaying the decoded channels along with the quality information on a monitor.

7.1.1 Functional diagram

Figure 6 shows the NET-MOZAIC process flow. The input signal originates from various IP multi-cast sources, depending on the application. Examples include:

- Satellite video encoders;
- DVB-C cable TV;
- DVB-CS satellite TV;
- DVB-T off-the-air TV broadcast; and
- VOD servers.

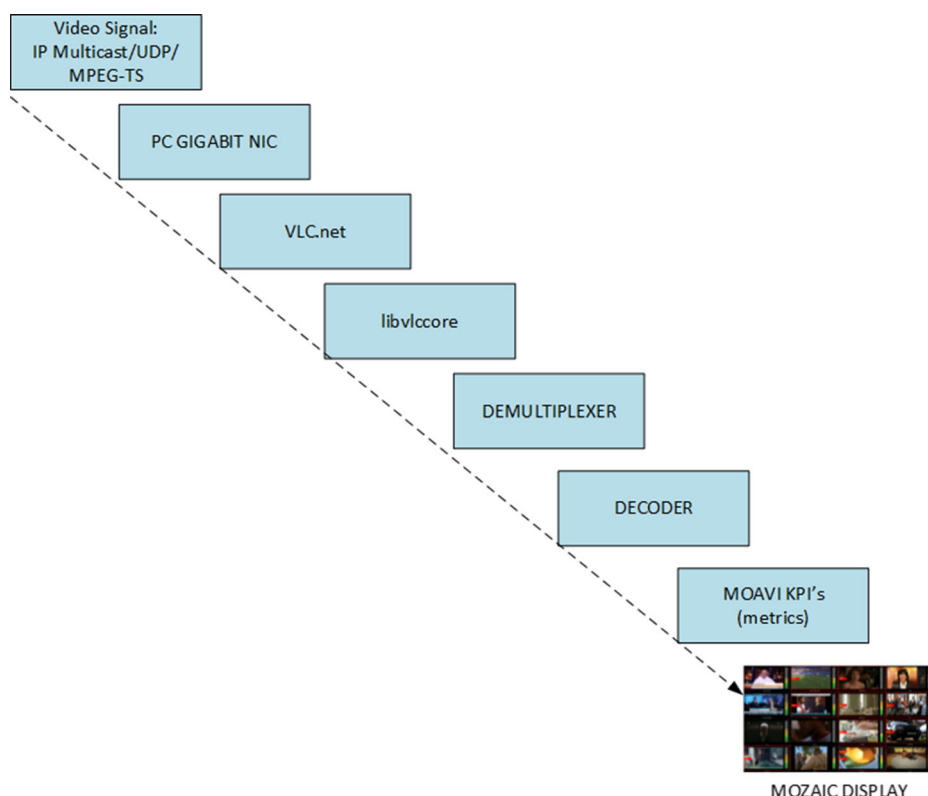


Fig. 6 Functional diagram of NET-MOZAIC

The NET-MOZAIC is a Windows application that runs on high performance PC. The higher the performance of processor and graphics coprocessor the higher the number of channels that can be simultaneously decoded and analysed for content quality. In our tests, a PC with XEON quad-core 5560 processor was used. Best performance achieved was 16 channels (one HD and 15 SD).

A typical IPTV operator offers a minimum 100 TV channels and typically 200-500 TV channels. To analyse this many channels a round robin method is used, with a limited time window per group of channels. The components of NET-MOZAIC application are shown in the Fig. 6 in an order of use.

7.1.2 Key features

The IPTV packets enter the NET-MOZAIC probe via Network Interface Card Giga-bit Ethernet. Each TV channel is encapsulated typically in UDP/MPEG-TS. Typical encoding carried in the MPEG-TS frames is MPEG-4 or in diminishing MPEG-2 format. Other less popular encoding standards are also possible.

The key features of NET-MOZAIC are:

- Simultaneous full motion preview and image analysis of up to 16 SD and/or HD channels;
- Round-robin for up to hundreds of IPTV channels;

- Integrated with the Net-xTVMS system for centralized control/access or stand alone operation;
- Support to MPEG-2 and H.264/AVC;
- Support to audio codecs AC-3, MPEG-1 Level 2, MPEG-2 AAC, MPEG-4 AAC;
- Support to UDP or UDP/RTP encapsulation;
- All metrics have user defined alarm thresholds;
- Long term storage of metrics; and
- Optional HDMI/IP encoding of a 16 channel group for advertising.

As a result of ongoing collaboration with AGH MOAVI research team the NET-200 is being further developed with addition of new metrics for image analysis. Periodic new software releases will be announced.

7.1.3 MOAVI KPI metrics implemented in the NET-MOZAIC

The MOAVI KPI metrics implemented in the NET-MOZAIC are:

- Blur;
- Exposure time (brightness problem);
- Noise (within video picture);
- Block loss;
- Blocking; and
- Freezing (frozen picture).

In addition several non-MOAVI metrics were added, which include:

- Black screen;
- Contrast problems;
- Flickering;
- Audio silence detection;
- Audio clipping detection;
- No audio; and
- No video.

The metrics algorithms are handled by VLC plug-ins. Each metric triggers a yellow or red QoE alarm when the distortion/disturbance level exceeds the predefined thresholds. These thresholds are user-defined as shown in the set-up menu displayed in Fig. 7. For each of the 16 channels, the alarms are displayed as color (yellow or red) frames around the channel window. The name of the metric causing the alarm appears as text on the red or yellow bar, as shown in Fig. 8.

The audio signal level for each channel is shown as a single or dual (stereo) color bar. The NET-MOZAIC layout can be adjusted to customer preference with more space for specific channels, typically for HD channels. For an example, see Fig. 9.

7.2 NET-MOZAIC in the NET-xTVMS system

The NET-MOZAIC can be used as a stand-alone QoE system for content analysis or it can be a part of an integrated NET-xTVMS system, as shown in the example in Fig. 10. This NET-xTVMS QoS/QoE monitoring system has two parts: the head-end and the network.

The head-end focuses multi-cast TV programming from multiple sources, such as satellite digital TV, video-on-demand (VoD) and off-the-air TV. This programming is transported to the head-end unencrypted. The NET-MOZAIC probe is also located at the head-end

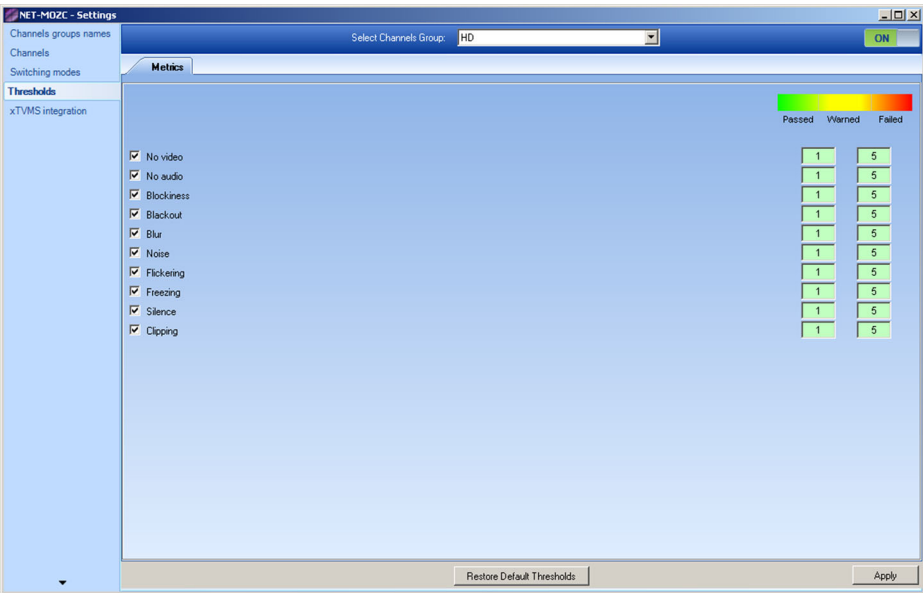


Fig. 7 Set up menu of NET-MOZAIC

and connected via Ethernet links to all IPTV channels in the unencrypted format. Next, the channels are encrypted and sent out to the core and edge routers. Finally, these encrypted channels are delivered to the customer set top boxes, where the programming gets decrypted. Therefore the NET-MOZAIC analyses image quality prior to the transmission throughout the IP network.



Fig. 8 Example of screens of NET-MOZAIC showing the alarms. Alarms are displayed as color (yellow or red) frames around the channel window. The name of the metric causing the alarm appears as text on the red or yellow bar

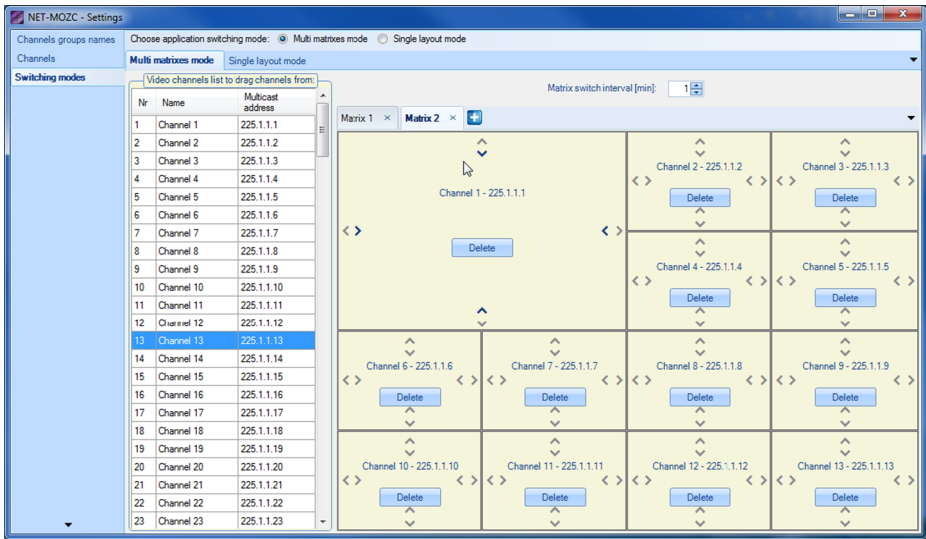


Fig. 9 NET-MOZAIC layout settings, which can be adjusted to customer preference with more space for specific channels

The network part of NET-xTVMS uses different probes, called NET-200 that perform QoE/QoS analysis on IP transport, specifically on MPEG-TS packets. The NET-200 probes are located typically at the hubs, near core and/or edge routers. Over 40 transport metrics are used to analyse MPEG-TS packet losses, jitter, transport quality index, TR-101290, etc.

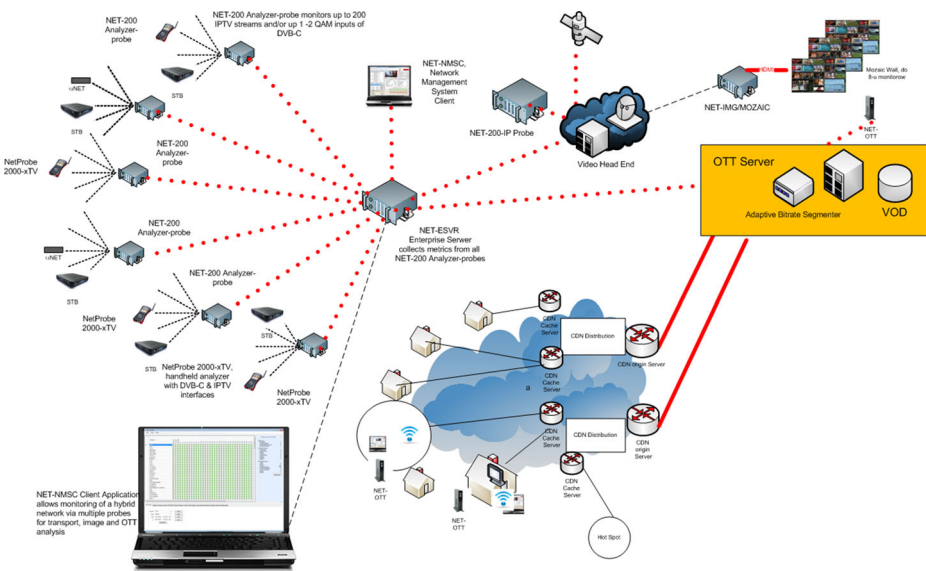


Fig. 10 Integration of NET-MOZAIC in the NET-xTVMS System

8 Conclusions and future work

In addition to traditional QoS, QoE poses a challenge for Internet service providers, audiovisual services, broadcasters, and new OTT services. Therefore, objective audiovisual metrics are frequently dedicated to monitoring, troubleshooting, investigating, and setting benchmarks of content applications working in real-time or off-line. The concept proposed here, MOAVI, is able to isolate and focus investigation, set-up algorithms, increase the monitoring period, and guarantee better prediction. MOAVI artefact KPIs are classified as four directories based on their origins: capturing, processing, transmission, and displaying. In the paper, we presented experiments carried out over several steps using four experimental set-ups for concept verification. The methodology took into account annoyance and visibility thresholds. The experimental methodology was adapted from ITU-T Recommendations: P.800, P.910 and P.930. We also presented the results of KPI verification tests. Finally, we described the first implementation of MOAVI KPI in a commercial product, the NET-MOZAIC probe. The probe is offered by Net Research, LLC as part of NET-xTVMS IPTV and CATV monitoring systems.

In terms of future work, we plan to provide more tools and evaluation methods for validating the visual content from the perspective of QoS and QoE. The objective will be to study and develop new methods and algorithms to deploy distributed probes for measuring (in real-time) the perceptual quality of video, considering content and context of information for multimedia entertainment and security scenarios. Finally, the work will target proposals for ITU Recommendations.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

1. Cerqueira E, Janowski L, Leszczuk M, Papir Z, Romaniak P (2009) Video artifacts assessment for live mobile streaming applications. In: Mauthe A, Zeadally S, Cerqueira E, Curado M (eds) Future multimedia networking. Lecture notes in computer science, vol 5630. Springer, Berlin / Heidelberg, pp 242–247. doi:[10.1007/978-3-642-02472-6_26](https://doi.org/10.1007/978-3-642-02472-6_26)
2. Farias MCQ, Heynderickx I, Macchiavello Espinoza BL, Redi JA (2013) Visual artifacts interference understanding and modeling (VARIUM). In: Seventh international workshop on video processing and quality metrics for consumer electronics, vol 1. Scottsdale
3. Farias MCQ, Mitra SK (2012) Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance. *J Electron Imaging* 21(4):043013–043013. doi:[10.1117/1.JEI.21.4.043013](https://doi.org/10.1117/1.JEI.21.4.043013)
4. Gustafsson J, Heikkilä G, Pettersson M (2008) Measuring multimedia quality in mobile networks with an objective parametric model. In: 15th IEEE international conference on image processing, 2008. ICIP 2008. pp 405–408. doi:[10.1109/ICIP.2008.4711777](https://doi.org/10.1109/ICIP.2008.4711777)
5. International Telecommunication Union: ITU-T P.800, Methods for subjective determination of transmission quality (1996). <http://www.itu.int/rec/T-REC-P.800-199608-I>
6. International Telecommunication Union: ITU-T P.930, Principles of a reference impairment system for video (1996). <http://www.itu.int/rec/T-REC-P.930-199608-I>
7. International Telecommunication Union: ITU-T P.910, Subjective video quality assessment methods for multimedia applications (1999). <http://www.itu.int/rec/T-REC-P.910-200804-I>
8. International Telecommunication Union: ITU-T J.144, Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference (2004). <http://www.itu.int/rec/T-REC-J.144-200403-I>

9. International Telecommunication Union: ITU-T J.246, Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference (2008). <http://www.itu.int/rec/T-REC-J.246-200808-I>
10. International Telecommunication Union: ITU-T J.247, Objective perceptual multimedia video quality measurement in the presence of a full reference (2008). <http://www.itu.int/rec/T-REC-J.247-200808-I>
11. International Telecommunication Union: ITU-R BT.500-12, Methodology for the subjective assessment of the quality of television pictures (2009). <http://www.itu.int/rec/R-REC-BT.500-12-200909-I>
12. International Telecommunication Union: ITU-T J.249, Perceptual video quality measurement techniques for digital cable television in the presence of a reduced reference (2010). <http://www.itu.int/rec/T-REC-J.249-201001-I>
13. International Telecommunication Union: ITU-T J.341, Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference (2011). <http://www.itu.int/rec/T-REC-J.341-201101-I>
14. International Telecommunication Union: ITU-T J.342, Objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference signal (2011). <http://www.itu.int/rec/T-REC-J.342-201104-I>
15. Karunasekera SA, Kingsbury NG (1995) A distortion measure for blocking artifacts in images based on human visual sensitivity. *Image Processing. IEEE Trans* 4(6):713–724
16. van Kester S, Xiao T, Kooij RE, Brunnström K, Ahmed OK (2011) Estimating the impact of single and multiple freezes on video quality, vol 7865, pp 786500–786500–10. doi:10.1117/12.873390
17. Leszczuk M, Hanusiak M, Blanco I, Wyckens E, Borer S (2013) Key indicators for monitoring of audiovisual quality. In: Cetin EA, Salerno E (eds) MUSCLE international workshop on computational intelligence for multimedia understanding. ERCIM, IEEE Xplore, Antalya, Turkey, p 6
18. Punchihewa A, Bailey DG (2002) Artefacts in Image and Video Systems: classification and mitigation. In: *Image and vision computing New Zealand*
19. Rohaly AM, Corriveau P, Libert J, Webster A, Baroncini V, Beerends J, Blin JL (2000) Video quality experts group: current results and future directions
20. Romaniak P, Janowski L, Leszczuk M, Papir Z (2011) A no reference metric for the quality assessment of videos affected by exposure distortion. In: 2011 IEEE international conference on multimedia and expo (ICME), pp 1–6
21. Romaniak P, Janowski L, Leszczuk M, Papir Z (2012) Perceptual quality assessment for h.264/avc compression. In: Consumer communications and networking conference (CCNC), 2012. IEEE, pp 597–602
22. Takahashi A, Schmidmer C, Lee C, Speranza F, Okamoto J, Brunnström K, Janowski L, Barkowsky M, Pinson M, Staelens Nicolas Huynh Thu Q, Green R, Bitto R, Renaud R, Borer S, Kawano T, Baroncini V, Dhondt Y (2010) Report on the validation of video quality models for high definition video content. Tech. rep., Video Quality Experts Group
23. Takahashi A, Yamagishi K, Kawaguti G (2008) Global standardization activities recent activities of qos / qoe standardization in itu-t sg12, vol 6, pp 1–5
24. Wyckens E (2011) Proposal studies on new video metrics. In: Webster A (ed) Hillsboro Meeting, V Q E G 17, Orange Labs, Video Quality Experts Group (VQEG), Hillsboro



Mikołaj Leszczuk's professional career started already in the period of study in ComArch SA as the Manager of the Multimedia Technology Department, since 1999, an employee of the AGH Department of Telecommunications. His early responsibilities included design and implementation of digital video libraries. In 2000 he moved from Poland to a four-month scholarship at the Universidad Carlos III de Madrid (Madrid, Spain), where he prepared a thesis ("Implementation of the Recognition Engine's Search Module in Multimedia Application"). After returning to Poland he was employed in the Department of Telecommunications as a research and teaching assistant, and in 2006, past the defense of doctoral dissertation ("Analysis of the Possibilities of Building a Web Application Access Digital Video Libraries"), as an assistant professor. His current research interests are focused on broad processing systems and multimedia data analysis, with particular emphasis on Quality of Experience. He is the author of over 100 scientific publications, 19 of which are published in the journals of the JCR database. He participated or participates in more than 20 major research projects of FP4, FP5, FP6, FP7, OPIE, Culture 2000, PHARE, eContent+, Eureka! and others. In the years 2009-2014, he was the administrator of a large, international INDECT research project, thanks to which, a set of solutions for intelligent surveillance and automatic detection of suspicious behavior and violence in an urban environment has been created. He is a member of VQEG (Video Quality Experts Group, a member of the board), IEEE (Institute of Electrical and Electronics Engineers), and GAMA (Gateway to Archives of Media Art). It works with the VQiPS (Video Quality in Public Safety) organization. More information: <http://www.linkedin.com/in/miklesz>.



Mateusz Hanusiak received the M.S. degree in Electronics and Telecommunication in 2014 from AGH University of Science and Technology in Cracow. In 2012-2014 he worked on Project INDECT in the field of video quality assessment.



Mylène C. Q. Farias received her B.Sc. degree in electrical engineering from Universidade Federal de Pernambuco (UFPE), Brazil, in 1995 and her M.Sc. degree in electrical engineering from the Universidade Estadual de Campinas (UNICAMP), Brazil, in 1998. She received her Ph.D. in electrical and computer engineering from the University of California Santa Barbara, USA, in 2004 for work in no-reference video quality metrics. Dr. Farias has worked as a research engineer at CPqD (Brazil) in video quality assessment and validation of video quality metrics. She has also worked for Philips Research Laboratories (The Netherlands) in video quality assessment of sharpness algorithms and for Intel Corporation (Phoenix, USA) developing no-reference video quality metrics. Currently, she is an Assistant Professor in the Department of Electrical Engineering at the University of Brasilia (UnB), where she is a member of the Graduate Program in Informatics and of the Graduate Program on Electronic Systems and Automation Engineering (PGEA). Dr. Farias is a researcher of the Group of Digital Signal Processing of UnB and her current interests include video quality metrics, video processing, multimedia signal processing, watermarking, and information theory. Dr. Farias is a member of IEEE and of the IEEE Signal Processing Society.



Emmanuel Wyckens research engineer for operational audiovisual services, graduated in Electronic Engineering at Valenciennes University (France), specializing in the design of audiovisual broadcasting solutions.

In 2000 he joined Orange, being involved in the improvement of MPEG-2/4 codec video quality and the optimization of audiovisual chain tools. Mr Wyckens current activities was in the field of subjective, objective audiovisual quality in multimedia applications, and for high definition television domains. He participated in the work of EBU/ITU projects for standardizing subjective methodologies. Recently, he has changed for involving in audiovisual design services focus in over the top context.



George Heston 11 years of professional experience as an electronics engineer and principal engineer at Comsat labs, American Satellite, Fairchild Industries, Logicon, AMF Electrical Products Div. Developed products for satellite television scrambling receiver, voice compression, data encryption, energy monitoring and others. Co-founder and president of Compresion Techniques Corporation, a telecom test equipment manufacturer known for the innovative T-STAR product line. Founder and president of Net Research Corporation with subsidiary R&D company in Poland. Developed NetProbe 1000&2000 telecom analyzers. Member of FETEST, LLC management team. Chief Architect – designed and market IPTV/triple play handheld analyzer, software based feStream Expert IPTV analyser and feStreamXMS IPTV monitoring system. Education: M.S.E.E, Lodz Polytechnic; M.S. Computer Science, George Washington University.