On the Subjective Assessment of the Perceived Quality of Medical Images and Videos

Lucie Lévêque, Hantao Liu
School of Computer Science and Informatics
Cardiff University
Cardiff, United Kingdom
LevequeL@cardiff.ac.uk

Sabina Baraković[1], Jasmina Baraković Husić[2]
Faculty of Traffic and Communications[1]
Faculty of Electrical Engineering[2]
University of Sarajevo
Sarajevo, Bosnia and Herzegovina

Asli Kumcu, Ljiljana Platisa
imec-TELIN-IPI
Ghent University
Ghent, Belgium

Maria Martini
School of Computer Science and Mathematics
Kingston University
London, United Kingdom

Rafael Rodrigues*, António Pinheiro*
Instituto de Telecomunicações
And Universidade da Beira Interior
Covilhã, Portugal

Meriem Outtas, Lu Zhang
Industrial Computer Science and Electronics Group
National Institute of Applied Sciences (INSA) of Rennes
Rennes, France

Athanassios Skodras
Faculty of Engineering
University of Patras
Patras, Greece

***Abstract*—Medical professionals are viewing an increasing number of images and videos in their clinical routine. However, various types of distortions can affect medical imaging data, and therefore impact the viewers' experienced quality and their clinical practice. Thus it is necessary to quantify this impact and understand how the viewers, i.e., medical experts, perceive the quality of (distorted) images and videos. In this paper, we present an up-to-date review of the methodologies used in the literature for the subjective quality assessment of medical images and videos and discuss their merits and drawbacks depending on the use case.**

Keywords—medical image quality; image quality assessment; video quality assessment; subjective experiment; MOS

I. Introduction

Medical imaging provides clinical information either unavailable by other means or with reduced invasiveness, playing a key role both in diagnosis and adequate treatment planning and monitoring. With about one billion exams performed per year [1], radiology encompasses a wide variety of imaging modalities (e.g. planar radiography and Computed Tomography (CT) scans [2], ultrasound [3]-[4], mammograms [5], Magnetic Resonance Imaging (MRI) [6]-[8], Positron Emission Tomography (PET) scans [9]), the use of which is complementary to other medical specialties. Both 2D and 3D content may be generated from these modalities, as well as video content. Besides radiology, other imaging modalities are commonly applied in diagnosis and treatment planning, such as pathology slides [10] and endoscopic surveys [11]-[13]. Furthermore, with the advances in telemedicine, and particularly in image-guided surgery and tele-surgery [14]-[15], image and video are also utilised in real-time frameworks. Therefore, large amounts of visual contents are continuously created for viewing or manipulation by medical professionals in their routine practice.

According to [16], the quality of experience (QoE) refers to the degree of delight or annoyance of the user of an application or a service; it involves both subjective (viewer's comfort) and objective (clinical task success) components. Despite the continuous evolution in how visual contents are acquired, stored, accessed and viewed, distortions or artifacts may be induced in image or video in one or several of the following steps: acquisition, processing, compression, transmission, and display. Such distortions or artifacts, together with the viewing conditions (e.g., room lighting, viewing distance), may affect the perceived image quality, but also compromise the diagnostic quality of the content [13]. As a consequence, exam interpretation may be hindered, leading to medical errors [17]. Thus, in the case of medical image and video, both perceived and clinical quality affect the QoE of medical professionals.

To minimise potential diagnostic errors caused by visual distortions or to achieve a satisfactory quality for image-guided

*This work was funded by FCT-Portugal under the project UID/EEA/50008/2013 and the research grant SFRH/BD/130858/2017.

978-1-5386-2605-4/18/$31.00 ©2018 IEEE

surgery, it is necessary to understand how the medical professionals perceive the quality of distorted medical images and videos. Psychovisual experiments have been conducted for various medical applications to learn how distinct distortions affect the perceptual quality in medical imaging, as this latter is critical to clinical practice; that is to say, how the quality can fulfil the medical professionals' expectations and help them in their routine practice. In recent years, medical image perception research has been growing, on one hand, to evaluate the perception of specific imaging distortions, and on the other hand, to consider imaging modalities outside of traditional radiology, such as pathology and surgical video. In both of these contexts, subjective quality assessment may contribute to the definition of default conditions and parameters, in terms of image acquisition, image/video encoding and content display, to assure adequate QoE in various medical image applications.

In this paper, a review of the methodologies used in the literature to assess the perceived quality of medical contents is conducted, illustrated by representative example studies. The studies chosen as examples are relatively recent (i.e., not older than 10 years old) and represent diverse modalities and contexts of application. Moreover, we discuss the presented subjective assessment methodologies and make recommendations for future studies in medical image quality assessment with respect to the selection of subjective test methodologies, the data analyses to be carried out, and consider the effect of other influential factors, such as the content type (image or video), the image modality and the expertise of test subjects.

II. Review of the Methodologies Used

Various methodologies offered by the International Telecommunication Union (ITU) have been used to assess the perceived quality of medical content. In this section, we present an overview of the methods and studies in the literature (see Table I). The existing methods can be divided into two groups: single stimulus (SS) and multi stimulus (MS). The stimulus – a 2D image or short video consisting of medical content – is presented to the observer, who then scores the perceived quality of the stimulus of interest using a discrete or continuous scale, typically containing five descriptors: *Bad*, *Poor*, *Fair*, *Good*, and *Excellent* quality. In SS methods, one stimulus is presented and scored. In MS methods, two or more stimuli are presented for reference but only the stimulus of interest is scored.

A. Single Stimulus Methods

Kara et al. [18] used the Absolute Category Rating (ACR) method [19] with to study the effects of angular resolution and light field reconstruction of 3D heart images. The ACR method is a SS method where the stimulus is evaluated on a discrete quality scale. Kara et al. chose a 10-point scale for their tests and recruited 20 observers, 8 were medical experts and 12 non-experts. Based on the regression analysis, the results identified a breakpoint of excellence at 75 views and showed that observers were more sensitive to degradations in texture than to a fewer number of views.

Platisa et al. [10] investigated the effects of blurring, colour, gamma parameters, noise, and image compression on animal digital pathology images (dog gastric fundic glands and foal liver). They conducted an image assessment study with 6 veterinary pathologists, 7 veterinary students, and 11 imaging experts using the Single Stimulus Hidden Reference Removal (SS-HRR) method [21] with a 6-point ACR scale. Using median opinion scores and Kruskal-Wallis non-parametric one-way Analysis of Variance (ANOVA), they observed disagreement between the quality ratings made by different expertise groups, warning against the use of psycho-visual responses of subjects who are not experts in pathology to guide the development of any pathology-specific image algorithms or imaging systems.

Tulu et al. [20] studied the effects of delay, jitter, and packet loss ratio on ophthalmology videos in the context of telemedicine, using the Single Stimulus Continuous Scale (SSCQS) method [22]. The configurations are similar to the ones of the ACR method, but a continuous scale was used in this case (i.e., a 100-point scale). Using ANOVA, they found that the perceived quality depended not only on technical parameters such as jitter and delay, but also on the transmission success of critical frames, i.e., the frames which are used to make a diagnosis, as they play a decisive role for the viewer.

B. Multi Stimulus Methods

Suad et al. [8] and Chaabouni et al. [11], respectively, used the Double-Stimulus Impairment Scale (DSIS) [19] and the Double-Stimulus Continuous Quality Scale (DSCQS) [22] for their experiments. These two DS methods employ a similar presentation protocol. With the DSIS method, the reference stimulus is presented first, followed by the distorted stimulus. For the DSCQS method, the reference and distorted stimuli are presented in random order and this presentation is repeated a second time. Suad et al. used the DSIS method to analyse the impact of different common types of distortion (i.e., additive Gaussian noise, blurring, JPEG compression, salt and pepper and sharpness) on brain MR images. A group of 15 doctors participated in the study, where they were asked to evaluate the quality of the images. The results showed that the perceived quality is strongly affected by the distortions, with the highest quality ratio for the sharpness and the poorest quality given by Gaussian noise. Chaabouni et al. made use of the DSCQS method on laparoscopic surgery videos to evaluate the impact of H.264 compression and analysed their data with some correlation metrics and a regression analysis. They found that compression artifacts could be noticeable for compression ratios of 100:1 up to 270:1.

Another study on H.264 encoded laparoscopic videos was conducted by Münzer et al. [12]. A group of 37 medical experts participated in a double session test, using the DSCQS method to evaluate the impact of resolution and the constant rate factor (CRF) changes on overall image and semantic quality. The results suggested that an acceptable quality may be achieved even reducing resolution down to 640×360 and with CRF = 26. With this set-up, storage requirements would drop to 12.5% of current practice.

Two studies on compressed videos were carried out by Razaak et al. [4] and by Usman et al. [13]. Both assessed the impact of the quantization parameter (QP) on both visual and diagnostic quality of HEVC compressed videos, using the DSCQS method. The first study addressed ultrasound video, the second one video from wireless capsule endoscopy. A total of 20 observers participated in the first study (16 non-experts and 4 medical specialists), 25 observers in the second one (19 non-experts and 6 medical experts). Experimental results, analysed with correlation metrics, recommended QP threshold values below 35 for acceptable diagnostic quality in the first case and of 35 and 37 in the second in order to provide satisfactory diagnostic and visual quality, respectively.

Nouri et al. [15] also made use of the DSCQS method on 4 videos representing different stages of a laparoscopic surgery (a type of minimally invasive surgery). Using a regression analysis, the authors found a quality threshold at bit rate = 3.2 megabits per second (or compression ratio 90:1 for MPEG2 compression), below which the surgeons considered the perceived image quality as too poor to perform the tasks.

TABLE I. REVIEW OF THE METHODOLOGIES USED IN THE LITERATURE FOR THE SUBJECTIVE ASSESSMENT OF MEDICAL CONTENT.

Content type	**Source article**	**Independent variables (# levels)**	**Method**	**Scale**[1]	**Participants**	**Statistical analysis**	**Scoring**	**Key findings**
MR images	Chow et al. [6]	Distortion type (6) Distortion level (5)	SDSCE	C	28 medical observers	T-test Correlations Regression analysis	Overall quality	Perceived quality of MR images is strongly affected by Rician Noise and Gaussian White Noise
MR images	Liu et al. [7]	Image modality and content Distortion type (7) Energy (5)	SDSCE	C	18 medical expert observers	ANOVA	Overall quality	MRI content influences the relative impact of artifacts. Distortions with a flat spectral power density are more annoying
MR images	Suad et al. [8]	Distortion type (5)	DSIS	D (5-pt)	15 medical observers	Not specified	Overall quality	Different types of distortions and noise affect the perceived quality
Pathology images	Platisa et al. [10]	Scene (12) Distortion type	ACR	D (6-pt)	6 medical expert, 7 trainee and 11 non-expert observers	Median opinion score ANOVA	Overall quality, blur and noise disturbance, contrast, colour saturation	Different observer profiles differently rate perceived image quality
3D heart images	Kara et al. [18]	Angular resolution Light field reconstruction	ACR	D (10-pt)	8 medical and 12 non-medical observers	Regression analysis	Overall quality	Observers are more sensitive to degradations in texture than to lower number of views
Endoscopic videos	Chaabouni et al. [11]	Scene (4) Bit rate (11)	DSCQS	C	14 medical observers	Correlations Regression analysis	Overall quality	Video could be lossy encoded up to a determined threshold while maintaining observer satisfactions
Endoscopic video	Munzer et al. [12]	Scene Resolution (4) Constant rate factor (6)	DSCQS	C	37 medical expert observers	Not specified	Overall quality, semantic quality	Laparoscopic videos may be further compressed than current practice maintaining overall image and semantic quality
Endoscopic videos	Usman et al. [13]	Scene (10) Quantization parameter (8)	DSCQS	C	19 non-expert and 6 medical expert observers	Correlations	Overall quality, diagnostic capability	A maximum value for the HEVC quantization parameter was recommended for endoscopic video
Tele-surgery video	Nouri et al. [15]	Scene (4) Bit rate (5-8)	DSCQS	C	7 medical expert observers	Regression analysis	Overall quality	Lossy compression could be used up to a certain threshold while preserving perceived visual quality
Ultrasound videos	Razaak at al. [4]	Scene (9) Bit rate (8)	DSCQS – type II	C	4 medical and 16 non-medical observers	Regression analysis Correlations	Overall quality, diagnostic quality	HEVC compression threshold keeping diagnostic accuracy identified for different types of content
Ultrasound videos	Lévêque et al. [3]	Scene (4) Encryption (2) Bit rate (7)	SAMVIQ	C	17 medical observers	ANOVA	Overall quality	The impact of visual content and compression configuration on the perceived quality of videos is found to be significant
Ophthalmo-logy videos	Tulu et al. [20]	Delay (5) Jitter (5) Packet loss ratio	SSCQS	C	15 medical observers	ANOVA	Overall quality, clinical decision-making capability	Subjective quality not only depends on jitter and delay, but also on which critical frames the viewer is able to see and work with

[1] C: continuous, D: discrete

Chow et al. [6] carried out subjective experiments to assess the quality of MR images of the human brain, spine, knee and abdomen distorted with 6 types of distortion (Rician noise, Gaussian white noise, Gaussian blur, discrete cosine transform (DCT), JPEG compression and JPEG 2000 compression) at 5 different levels. They made use of the Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) methodology [21], where the reference image and its distorted version are displayed side by side on a monitor. The observers rated the distorted image by judging the differences with the original image. According to the results obtained by t-test, correlation and regression analysis, they declared that Rician and Gaussian noise strongly affected the quality of MR images.

The impact of a set of common distortions on the perceived quality of brain, liver, breast, foetus, hip, knee, and spine MR images was also studied by Liu et al. [7]. Ghosting, edge ghosting, white noise and coloured noise artifacts were simulated on MR scans with different content and acquisition protocols, for two experiments. The first experiment was divided in two parts: the first one included ghosting and white noise artifacts, while the second one included edge ghosting and coloured noise. Five different energy levels were defined for each artifact. In each part, a total of 30 stimuli were shown to 15 and 17 expert participants, respectively, using the SDSCE method with a scale from 0 to 100. For the second experiment, a similar procedure was followed with 18 expert subjects, using the 2 higher energy levels of all artifacts plus 3 new variations of the coloured noise, in a total of 112 stimuli. The scores obtained from a one-way ANOVA indicated that artifacts with a flat spectral power density were nearly twice as annoying as artifacts with a spectral power density similar to the original image, at the same energy level. The study also concluded that differences in content affect artifact visibility and impact.

Finally, Lévêque et al. [3] chose the Subjective Assessment Methodology for Video Quality (SAMVIQ) [23] to investigate the effect of H.264 and HEVC compression on hepatic ultrasound videos. SAMVIQ offers the visualisation of a short video through a graphic interface, where the observers can navigate among the reference and the distorted versions of a content. Two expertise groups participated in the tests: 8 radiologists and 9 sonographers. Their ANOVA results showed the impact of content and compression configuration (i.e., video codec and bit rate) on the perceived quality, as well as the impact of the specialty settings, as the sonographers were more annoyed by the distortions than the radiologists.

III. Discussion of the Existing Approaches

In this section, we elaborate more on the methodologies used in the literature and provide recommendations for researchers in the medical image perception field.

A. Single Stimulus Versus Multi-Stimulus Methods

As we noticed from the studies presented in the previous section, both single- and multi- stimulus methods can be used for the subjective assessment of perceptual quality of medical images and videos that differ in acquisition modalities, such as endoscopic [11]-[13], ultrasound [3]-[4], pathology [10], ophthalmology [20], etc. Each methodology has its own advantages. Experiments conducted using single stimulus (SS) methods are usually quicker to conduct than with double stimulus (DS) methods, and they avoid potential vote inversions as only one stimulus is rated at a time [24]. However, SS experiments may lead to score drift over the course of a session [25]. New methods have therefore been defined to combine the combine the advantages of both approaches, such as the SAMVIQ method [23] which allows the observers to review the reference and re-evaluate their scores.

Depending on the content type, i.e., images or videos, some methodologies may be more appropriate than others, regarding the length of the experiments for instance. Indeed, a major problem in medical image and video quality assessment may be finding medical professionals who have enough time to carry out the experiments. Reference [26] gives the length of total assessment time for 4 widely used methods, from shortest to longest: ACR, DSIS, SAMVIQ and DSCQS.

B. Influence Factors

The perceptual quality of medical images and videos can be affected by different influential factors (IFs), which can be grouped into three categories [16, 27]: system IFs, context IFs, and human IFs. The methods used for subjective assessment of medical content usually take into account the system IFs, which refer to properties and characteristics that determine the technically produced quality of medical image and video. There are four types of system IFs: content-related IFs, media-related IFs, network-related IFs and device-related IFs. Since the content-related and media-related IFs are interlaced, they can be discussed jointly.

As shown in Table I, there are many research studies dealing with the content/media-related IFs (e.g., distortion type, scene, bit rate, resolution, encryption, etc.), whereas the network-related IFs (e.g., delay, jitter, packet loss, etc.) have received much less attention. Moreover, there is no evidence that the device-related IFs have been considered in subjective assessment of perceptual quality of medical contents. Based on the non-exhaustive literature review from Table I, one may notice that the usually considered content/media-related IFs are scene [3], [10]-[13] and distortion type [3], [6]-[8]. These, along with other content/media-related IFs, are manipulated at different levels in order to conduct statistical analysis to describe their impact on the perceptual quality of medical images and videos.

In subjective user studies, the perceived quality of medical contents may vary with the viewing environments which have different context IFs. They refer to any situational property of the environment of medical images and videos that have an impact on perceptual quality. The most important context IFs in subjective assessment of perceptual quality of medical contents are physical and temporal IFs. While many subjective medical QA studies have been conducted in different environments and under various conditions, the studies likely do not consider these as physical or task IFs; thus the experimental protocol and analysis do not explicitly include nor control for these IFs. Therefore, a deeper and more comprehensive analysis of context IFs is required to properly assess the perceptual quality of medical content [28], by varying IFs within the same experiment, such as: different applications (e.g., diagnosis, surgery, training, etc.), clinical factors (e.g., emergency care, lesion suitableness, etc.), requirements (e.g., real-time/offline, location, etc.), medical data (e.g., clinical information, anatomical, functional, physiological), acquisition modalities (e.g., ultrasound, X-Ray, MRI), or data types (images and videos (monochrome, colour)).

The perceptual quality of medical images and videos may also be affected by the human IFs, which refer to any properties or characteristics of the human user that influence his/her perception of quality. Human IFs can be generally classified into two categories, i.e., low-level processing IFs (e.g., physical, emotional and mental constitution of human, etc.) and high-level processing IFs (e.g., demographic and socio-economic background, etc.). Based on the literature review from Table 1, it can be seen that there are almost no subjective user studies considering the impact of the low-level processing human IFs (e.g., emotional state of the medical observer) on the perceptual quality of medical images and videos. However, an important high-level processing human IFs that is usually taken into consideration is the number and expertise of the subjects.

Table I shows that the number of assessors varies from 6 to 37 medical observers. As previously mentioned, the availability of medical experts can be a limiting factor in medical QA studies. The expertise domain of the assessors span: the radiologist (specialised in medical imaging and trained to interpret the scans to make a diagnosis), the specialist (specialised in diagnosis and treatment of a particular organ), the surgeon (specialised in surgery with more anatomical knowledge and its clinical relevance), and the radiographer (trained to perform imaging scans). Lévêque et al. [3] concluded that sonographers (radiographers who perform ultrasound exams) are more annoyed by the distortions than the radiologists. Chaabouni et al. [11] conducted subjective tests with 14 doctors specialised in ENT (Ear Nose and Throat) and concluded that the highly compressed medical videos presented to doctors do not modify their perception and opinion. Liu et al. [7] studied the impact of distortions on perceived image quality of the MR images helping 17 clinical scientists or application physicists with a principal role in the review of diagnostic image quality. Zhang et al. [29] asked 4 radiologists and 8 naïve observers to complete a task of the detection of abnormality. Radiologists had a better detectability of hyper-signals than naïve observers on the MR images.

In a study on the quality assessment of compressed laparoscopic videos, Kumcu et al. [30] asked surgeons and non-experts to rate the overall quality of 20 sequences. Their first observation was that the subjective median scores were correlated between experts and non-experts with a Spearman correlation score of 0.83. Their second remark was that the surgeons had an ability to appreciate the specific anatomical structures when assessing the quality, while non-experts were insensitive to the content when evaluating the effects of compression. The authors suggested that non-experts should not be used as surrogates of surgeons for the quality judgment. Similar conclusion was drawn by Platisa et al. [10], but for non-expert versus digital pathologists.

According to the above-mentioned experiments, we may conclude that the expertise of the assessors must be carefully considered during the preparation of the subjective experiment in medical context. Nevertheless, non-medical or naïve assessors could be involved, if non-prior medical knowledge is required for certain applications (such as a pre-assessment task or a supplementary test).

C. Statistical Analyses

Some observers may generate dubious scores during experiments, due to a misunderstanding of the instructions or to a lack of engagement in the task [31]. It is therefore recommended to use an outlier detection and subject exclusion procedure, as given in ITU-R recommendation BT.500-11 [21]. Five of the twelve studies explicitly mention this 2-step pre-processing method: Chow et al. [6], Lévêque et al. [3], Münzer et al. [12], Liu et al. [7], and Usman et al. [13]. No subject was rejected in the first three, while the last two each discarded one observer. Two studies used other methods: a graphical technique by Platisa et al. [10] and removal of extreme scores by Kara et al. [18]. Finally, the five other studies did not mention any outlier removal procedure.

Subjective experiments in medical imaging can have different requirements compared to natural scene experiments. In a medical context, it is particularly important to test whether participants are consistent in their quality scoring, as their years of experience in medicine may affect their perception of visual distortions [32]-[33]. Therefore, it may be necessary to divide the observers into groups depending on their experience and/or specialty, as it has been done in some of the studies in Table I. The criterion used to categorise the groups should be fixed prior to recruitment of participants. In addition, a common way to analyse the impact of the participants on scoring is to conduct an Analysis of Variance (ANOVA) on the scores. ANOVA is used to compare the means of two or more independent samples when assuming normality and homogeneity of the variance. As seen in Table I, three studies used an ANOVA for statistical analysis. Lévêque et al. [3] found no significant difference between the experts within each specialty group. Similarly, Tulu et al. [20] found no difference between the experts. Liu et al. [7] found a significant difference between observers and thus normalised the raw scores using z-scores [34].

Two other main analyses were used in the studied articles, i.e., correlation and regression analysis, for different objectives. Correlation analysis has mostly been used to evaluate the relationship between existing image quality metrics and the human scores obtained. Chow et al. [6], Chaabouni et al. [11], Razaak et al. [4], and Usman et al. [13] made use of the Pearson linear correlation coefficient (PLCC) and Spearman's rank correlation coefficient (SROCC) to assess some widely used metrics, such as PSNR, SSIM, UQI, VIF, etc [35]. On the other hand, Kara et al. [18] and Nouri et al. [15] used a regression curve to estimate the scores according to independent variables (e.g. bit rate), which enabled them to define quality thresholds.

IV. Conclusions

In this paper, we presented diverse methodologies used in the literature for the subjective assessment of medical contents and detailed their assets and drawbacks depending on the context of use. We suggest that future studies consider the following points of attention: some methodologies may be quicker and therefore better suited for experiments with busy medical specialists; studies should consider explicitly incorporating clinical influential factors as variables; the expertise level or speciality of the observers should be carefully selected given the aim of the study; observer outlier detection methods should be used and documented; and transformation of raw scores based on preliminary analysis should be considered.

References

[1] E. Krupinski, "Current perspectives in medical image perception", *Attention, Perception & Psychophysics*, vol. 72, pp. 1205-1217, 2010.

[2] H. Zhang, J. Huang, J. Ma, Z. Bian, Q. Feng, H. Lu, Z. Liang, and W. Chen, "Iterative reconstruction for X-ray computed tomography using prior-image induced nonlocal regularization", *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2367-2378, 2014.

[3] L. Lévêque, W. Zhang, P. Parker, and H. Liu, "The Impact of Specialty Settings on the Perceived Quality of Medical Ultrasound Video", *IEEE Access*, vol. 5, pp. 16998-17005, 2017.

[4] M. Razaak, M. Martini, and K. Savino, "A study on quality assessment for medical ultrasound video compressed via HEVC", *IEEE Journal of biomedical and health informatics*, vol. 18, pp. 1552-1559, 2014.

[5] S. Taplin, C. Rutter, C. Finder, M. Mandelson, F. Houn, and E. White, "Screening Mammography: Clinical Image Quality and Risk of Interval Breast Cancer", *American Journal of Roentgenology*, vol. 178, pp. 797-803, 2002.

[6] L. Chow, H. Rajagopal, R. Paramesran, and Alzheimer's Disease Neuroimaging Initiative, "Correlation between subjective and objective assessment of magnetic resonance (MR) images", *Magnetic Resonance Imaging*, vol. 34, pp. 820-831, 2016.

[7] H. Liu, J. Koonen, M. Fuderer, and I. Heynderickx, "The relative impact of ghosting and noise on the perceived quality of MR images", *IEEE Transactions on Image Processing*, vol. 25, pp. 3087-3098, 2016.

[8] J. Suad, and W. Jbara, "Subjective quality assessment of new medical image database", *International Journal of Computer Engineering and Technology*, vol. 4, pp. 155-164, 2013.

[9] J. Disselhorst, M. Brom, P. Laverman, C. Slump, O. Boerman, W. Oyen, M. Gotthardt and E. Visser, "Image-quality assessment for several positron emitters using the NEMA NU 4-2008 standards in the Siemens Inveon small-animal PET scanner", *Journal of nuclear medicine*, vol. 51, pp. 610-617, 2010.

[10] L. Platisa, L. Van Brantegem, Y. Vander Haeghen, C. Marchessoux, E. Vansteenkiste, and W. Philips, "Psycho-visual evaluation of image quality attributes in digital pathology slides viewed on a medical color LCD display", *Proceedings of SPIE Medical Imaging,* vol. 8676, 2013.

[11] A. Chaabouni, Y. Gaudeau, J. Lambert, J.-M. Moureaux, and P. Gallet, "Subjective and objective quality assessment for H.264 compressed medical video sequences", *Proceedings of the 4th International Conference in Image Processing,* pp. 1-5, 2014.

[12] B. Münzer, K. Schoeffmann, L. Böszörmenyi, J. Smulders, and J. Jakimowicz, "Investigation of the Impact of Compression on the Perceptual Quality of Laparoscopic Videos," *IEEE 27th International Symposium on Computer-Based Medical Systems*, pp. 153-158, 2014.

[13] M. A. Usman, M. R. Usman, and S. Shin, "Quality assessment for wireless capsule endoscopy videos compressed via HEVC: From diagnostic quality to visual perception", *Computers in Biology and Medicine*, vol. 91, pp. 112-134, 2017.

[14] L. Lévêque, W. Zhang, C. Cavaro-Ménard, P. Le Callet, and H. Liu, "Study of Video Quality Assessment for Telesurgery", *IEEE Access*, vol. 5, pp. 9990-9999, 2017.

[15] N. Nouri, D. Abraham, J. Moureaux, M. Dufaut, J. Hubert, and M. Perez, "Subjective MPEG2 compressed video quality assessment: Application to tele-surgery", *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 764-767, 2010.

[16] P. Le Callet, S. Möller, and A. Perkis, "Qualinet white paper on definitions of quality of experience". *European Network on Quality of Experience in Multimedia Systems and Services,* 2012.

[17] E. Krupinski, "Improving patient care through medical image perception research", *Policy Insights from the Behavioral and Brain Sciences*, vol. 2, pp. 645-659, 2008.

[18] P. Kara, P. Kovacs, S. Vagharshakyan, M. Martini, S. Imre, A. Barsi, and T. Balogh, "Perceptual Quality of Reconstructed Medical Images on Projection-Based Light Field Displays", *eHealth*, pp. 476-483, 2017.

[19] Recommendation ITU-T P.910, *Subjective video quality assessment for multimedia applications*, 2008.

[20] B. Tulu, and S. Chatterjee, "Internet-based telemedicine: An empirical investigation of objective and subjective video quality", *Decisions Support Systems*, vol. 45, pp. 681-696, 2008.

[21] Recommendation ITU-R BT.500-13, *Methodology for the subjective assessment of the quality of television pictures*, 2012.

[22] Recommendation ITU-R BT.500-11, *Methodology for the subjective assessment of the quality of television pictures*, 2002.

[23] Recommendation ITU-R BT.1788, *Methodology for the subjective assessment of video quality in multimedia applications*, 2007.

[24] A. Kumcu, K. Bombeke, L. Platisa, L. Jovanov, J. Van Looy, and W. Philips, "Performance of Four Subjective Video Quality Assessment Protocols and Impact of Different Rating Preprocessing and Analysis Methods", *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 48-63, 2017.

[25] M. Pinson, and S. Wolf, "Comparing subjective video quality testing methodologies", *Visual Communications and Image Processing*, vol. 5150, pp. 573-582, 2003.

[26] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video", *2nd IEEE International Workshop on QoMEX,* 2010.

[27] U. Reiter, K. Brunnström, and K. De Moor, *Quality of Experience: advanced concepts, applications and methods*, T-Lab Series in Telecommunication Services, p. 55-72, 2014.

[28] C. Cavaro-Ménard, L. Zhang-Ge, and P. Le Callet, "QoE for Telemedicine: Challenges and Trends," *SPIE 8856, Applications of Digital Image Processing XXXVI*, vol. 8856, 2013.

[29] L. Zhang, C. Cavaro-Ménard, P. Le Callet, and L. Cooper, "The effects of anatomical information and observer expertise on abnormality detection task.", *Proc. SPIE Medical Imaging*, vol. 7966, 2011.

[30] A. Kumcu, K. Bombeke, H. Chen, L. Jovanov, L. Platisa, H. Luong, J. Van Looy, Y. Van Nieuwenhove, P. Schelkens, and W. Philips, "Visual quality assessment of H.264/AVC compressed laparoscopic video", *Proceedings of SPIE*, vol. 9037, 2014.

[31] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment", *Computer Graphics forum,* vol. 31, pp. 2478-2491, 2012.

[32] C. Nodine, H. Kundel, S. Lauver, and L. Toto, "Nature of expertise in searching mammograms for breast masses", *Academic Radiology*, vol. 3, pp. 1000-10006, 1996.

[33] E. Krupinski, and R. Weinstein, "Changes in visual search patterns of pathology residents as they gain experience", *Proceedings of SPIE*, vol. 7966, 2011.

[34] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling," *Proceedings of SPIE*, vol. 2451, pp. 90–101, 1995.

[35] Z. Wang, H. Sheikh, and A. Bovik, "Objective Video Quality Assessment", *The Handbook of Video Databases: Design and Applications*, pp. 1041-1078, 2003.

View publication stats