Comparative study of the methodologies used for subjective medical image quality assessment

Lucie Lévêque, Meriem Outtas, Hantao Liu, Lu Zhang

IOP Publishing Physics in Medicine & Biology
Journal **XX** (XXXX) XXXXXX https://doi.org/XXXX/XXXX

Comparative study of the methodologies used for subjective medical image quality assessment

Lucie Lévêque[1], Meriem Outtas[2], Hantao Liu[3], and Lu Zhang[2]

[1] Nantes Laboratory of Digital Sciences (LS2N), University of Nantes, Nantes, France
[2] Department of Industrial Computer Science and Electronics, National Institute of Applied Sciences, Rennes, France
[3]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

E-mail: lucie.leveque@univ-nantes.fr

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Healthcare professionals have been increasingly viewing medical images and videos in their routine clinical practice, and this in a wide variety of environments. Both the perception and interpretation of medical visual information, across all branches of practice or medical specialties (e.g., diagnostic, therapeutic, or surgical medicine), career stages, and practice settings (e.g., emergency care), appear to be critical for patient care. However, medical images and videos are not self-explanatory and, therefore, need to be interpreted by humans, i.e., medical experts. In addition, various types of degradations and artifacts may appear during image acquisition or processing, and consequently affect medical imaging data. Such distortions tend to impact viewers' quality of experience, as well as their clinical practice. It is accordingly essential to better understand how medical experts perceive the quality of visual content. Thankfully, progress has been made in the recent literature towards such understanding. In this article, we present an up-to-date state of the art of relatively recent (i.e., not older than ten years old) existing studies on the subjective quality assessment of medical images and videos, as well as research works using task-based approaches. Furthermore, we discuss the merits and drawbacks of the methodologies used, and we provide recommendations about experimental designs and statistical processes to evaluate the perception of medical images and videos for future studies, which could then be used to optimise the visual experience of image readers in real clinical practice. Finally, we tackle the issue of the lack of available annotated medical image and video quality databases, which appear to be indispensable for the development of new dedicated objective metrics.

Keywords: medical imaging, image quality assessment, subjective experiment, task performance, objective metrics

1. Introduction

Medical imaging involves several scanning techniques to visualise the interior of the human body, along with a representation of the functions of some organs and tissues. Medical images and videos constitute a major part of the information used by healthcare professionals in their clinical routine practice in order to provide diagnostic decisions and further treatments. Indeed, medical imaging is nowadays ubiquitous among various specialties including, but not limited to, radiology, surgery, cardiology, oncology, and pathology [1]. Such non-invasive technique allows the

provision of critical information sometimes unavailable otherwise. As an order of magnitude, the number of imaging examinations performed per year in radiology approximates a billion [1]. Several modalities are comprised within the radiology discipline, such as magnetic resonance (MR) images, X-ray images, ultrasound, computed tomography (CT), positron emission tomography (PET) scans, and screening mammograms [2]-[4]. It should however be noted that, on top of radiology, other medical imaging modalities are widely employed, in particular image-guided surgery, pathology slides, and endoscopic surveys [5]-[6]. Furthermore, telemedicine, referring to the "use of information and communication technologies to provide and support clinical healthcare at a distance" [7], yielded a novel practice where medical images and videos are acquired, transferred, and stored for diagnosis and treatment planning [8]. The perception and interpretation of visual information, across all specialties, career stages, and practice settings, are consequently critical to patient care and safety.

However, medical images and videos are not self-explanatory, i.e., their conclusions are not always obvious, and therefore need to be interpreted by humans. Unfortunately, the latter can be prone to errors caused by the inherent limitations of the human visual system (HVS). The HVS is a part of the central nervous system enabling humans to see their environment [9]. Visual attention represents a powerful mechanism of the HVS, which helps the human brain to continuously minimise the overloading amount of input into a manageable flow of information, reflecting the current needs of the organism and the external demands [10]. A better understanding of the perceptual factors intrinsic to the interpretation of medical visual content would allow the improvement of patient care thanks to a decrease in the number of diagnostic errors.

In spite of the recent progresses made in imaging technology in medicine, visual signal distortions or artifacts may arise during the acquisition, processing, compression, enhancement, restoration, transmission, display, and even reproduction steps [11]. Such quality degradation, appearing at the acquisition or post-processing stage, may affect the perceptual quality of medical visual content and potentially impact the accurate and efficient interpretation of images [12]-[13]. It is consequently essential to understand how medical professionals perceive visual content, and to use such knowledge to develop new solutions to improve clinical practice: this is the essence of the science of medical image perception. Image quality assessment is therefore critical to control and maintain the perceived quality of medical visual content. As human observers are the ultimate receivers of visual information, subjective quality assessment is considered the most reliable approach in the medical field, where patients' safety is the priority. The International Telecommunication Union (ITU) established standardised methods for the subjective quality evaluation of image and video content [14]-[16]. These tests involve clinical specialists as observers, who perform specific diagnostic tasks.

In this paper, we investigate the methodologies used for subjective medical image and video quality assessment over the past decade, with a view to present a comprehensive literature review covering diverse medical specialties and applications. This review can be considered an extension of our previous work [17], with several significant contributions to the medical image perception field. More precisely, we examine the existing studies on the evaluation of perceptual quality in radiology in Section 2, we consider the works relating to the assessment of surgery and other modalities in Section 3, and we review the articles dealing with task-based approaches in Section 4. Finally, we thoroughly discuss the studies presented (i.e., methodologies, data analysis, and further) in Section 5.

2. Review of the perceptual-based approaches used for the assessment of radiological image quality

In this section, an overview of the research on medical image quality assessment in radiology present in the literature is exposed. Table I summarises these studies.

2.1. Magnetic resonance (MR) imaging

In 2013, Suad *et al.* released a new medical image quality database, composed of medical images and their associated mean opinion scores (MOS) [18]. More precisely, the authors aimed to evaluate the impact of diverse distortion types on magnetic resonance (MR) images. To do so, they first created a new dataset made of twenty brain MR images. The resolution of these reference images was 512×512 pixels. They then chose five different types of distortions and noise, as follows: additive Gaussian noise, blurring, lossy JPEG compression, salt and pepper (impulse noise), and sharpness. In practical routine, these artifacts commonly appear at diverse stages of MR image processing steps, such as image acquisition and compression (i.e., both types of noise), compression (i.e., JPEG compression), filtering (i.e., blurring), and enhancement (i.e., sharpness). As recommended by the International Telecommunication Union (ITU), Suad *et al*. used the double stimulus impairment scale (DSIS) for their experiment [16]. With this method, participants simultaneously look at both a distorted image and its reference, and are asked to rate the quality of the distorted one on a five-degree scale. Fifteen specialist doctors in e-diagnosis from Baghdad Central Hospital, Iraq, were recruited to participate in the experiment. As mentioned previously, MOS were computed for each of the one hundred images after test completion. Statistical analyses were carried out to scrutinise the impact of each distortion type on perceived quality. In general, sharpness showed the highest quality ratio,

followed by JPEG compression, blurring, and salt and pepper. Finally, additive Gaussian noise presented the poorest quality scores.

A couple of years later, Rajagopal *et al*. published their research work [19]. Their objective was to compare subjective and objective quality assessment for MR images. With a view to achieve such a goal, they selected ten MR images from the OsiriX DICOM Viewer MRI database [20]. All images are of good quality, in grayscale, and were normalised to (0,255). The authors applied four different types of distortions on these reference images, i.e., Rician noise, Gaussian white noise, Gaussian blur, and discrete cosine transform (DCT) compression. These artifacts often occur in MR images. For instance, Gaussian noise appears when the signal-to-noise ratio (SNR) is greater than two, whereas Rician noise appears when the SNR is lower than two [21]. Ten research scholars from the Electrical Engineering department of the University of Malaya, Malaysia, were involved in the subjective experiment, carried out in a controlled office environment. They were invited to rate the perceived quality of the two hundred distorted stimuli as compared to the ten associated references, using the simultaneous double stimulus for continuous evaluation (SDSCE) method [16]. The difference mean opinion scores (DMOS) were computed from the recorded data. Note a high DMOS value corresponds to a low image quality and vice versa. Statistical analyses ran on the DMOS showed poor perceived quality for high DCT compression rate and high standard deviation for Rician and Gaussian white noises, and Gaussian blur. This indicates that participants were able to differentiate distinct levels of distortions. Moreover, results demonstrated that the DMOS values did not deviate much through the different levels of Gaussian blur and DCT compression. This can be explained by the fact that Gaussian blur only causes small details to be visible, and DCT compression only yields little quality loss. Four full reference image quality assessment (FR-IQA) metrics were applied to the database: peak signal-to-noise ratio (PSNR) [22], structural similarity index measure (SSIM) [23], noise quality measure (NQM) [24], and visual information fidelity (VIF) [25]. In general, the DMOS were well correlated with these objective metrics, i.e., with correlation values between 0.89 and 0.95 for NQM, 0.76 and 0.96 for PSNR, 0.79 and 0.95 for SSIM, 0.81 and 0.95 for VIF.

In 2016, the previous research team extended their medical image quality database [2]. Indeed, Chow *et al*. published a dataset containing a total of 775 MR images. They chose twenty-five good quality original stimuli from two databases, the OsiriX DICOM Viewer MRI database [20], and the Alzheimer's Disease Neuroimaging Initiative (ADNI) MRI database [26]. The original MR images are T1 weighted, T2 weighted, or proton density. They represent images of brain, abdomen, spine, and knee. Six types of distortion, at five different levels, were applied to each image, namely: Rician noise, Gaussian white noise, Gaussian blur, discrete cosine transform, JPEG, and JPEG2000 compressions. These three compression types are commonly used to compress MR images [27]-[28]. Twenty-eight research scholars participated in the experiment, where they were asked to rate the quality of all the distorted images using the SDSCE methodology [16]. Before analysing the scores obtained, the authors conducted an outlier detection and subject rejection procedure. DMOS values were calculated using the remaining raw scores. Data analysis showed, as expected, that DMOS values increase with an increase of the noise's standard deviation. Low DCT compression rate, low JPEG compression quality, and high JPEG2000 compression rate resulted in high DMOS values, that is to say, poor image quality. Furthermore, a high correlation was found between the DMOS values and thirteen FR-IQA metrics studied, i.e., SNR [22], PSNR [22], SSIM [23], multi-scale SSIM (MS-SSIM) [29], feature similarity index measure (FSIM) [30], information fidelity criterion (IFC) [31], NQM [24], weighted SNR (WSNR) [24], VIF [25], VIF in pixel domain (VIFP) [25], universal image quality index (UQI) [32], information-weighted PSNR (IW-PSNR) [34], and information weighted SSIM (IW-SSIM) [33]. NQM presented the highest correlation (i.e., 0.94), whereas UQI showed a lowest correlation (i.e., 0.81). The authors have decided to release their database for the medical image perception field.

The same year, Liu *et al.* published their own study, where the goal was to measure how ghosting and noise impact MR images [34]. Two distinct experiments were conducted by the authors. The first one aimed to investigate the relative effect of structured vs. unstructured artifacts. For this part of the study, three original high-quality MR images were collected using Philips Achieva 1.5T system. Two represent brains, while the third one is a liver image. Each original image was distorted using four types of distortions, i.e., ghosting, edge ghosting, white noise, and coloured noise, at five different energy levels, for a total of sixty stimuli. The second part of the experiment was carried out as an extension of the first one. Indeed, eight reference images were used then; on top of the three previous images, stimuli of breast, hip, knee, and spine were added. The four distortion types presented previously were applied to the images, however, at only two energy levels. This resulted in a total of 112 stimuli. Fifteen, and eighteen clinical scientists and applications specialists from Philips Healthcare in The Netherlands were recruited for the first and second experiments, respectively. They were asked to rate the quality of the distorted stimuli with respect to the original stimuli using the simultaneous double stimulus (SDS) method [16]. After an outlier detection and subject exclusion procedure, the raw scores were calibrated and averaged towards MOS. Statistical analyses of the scores collected during the first phase of the experiment revealed the significant effects of image content, type of artifact, and

energy level on the perceived quality. In general, images distorted with coloured noise (i.e., unstructured artifacts) were scored higher than those with edge ghosting (i.e., structured artifacts). It can further be noted that images affected with ghosting were scored higher than those with white noise. As for images with coloured noise, they were scored higher than with edge ghosting. Results of the second part of the experiment showed that ghosting generally yielded higher scores than white noise, and that not all source images had the same overall quality. Furthermore, data demonstrated that edge ghosting deteriorated quality the most, followed by white noise, ghosting, and coloured noise. To conclude, structured artifacts deteriorated quality more than unstructured ones, and white artifacts more than coloured ones.

A few months ago, Mason *et al.* released a study comparing subjective and objective quality of MR images [35]. They selected eighteen reference images from their hospital's picture archiving and communication system (PACS), nine being from the abdomen and nine from the brain. More precisely, of the abdomen images, three were of liver, three of pancreas, and three of prostate. Six different degradations techniques were applied to the original stimuli, i.e., white Gaussian noise, Gaussian blurring, Rician noise, undersampling of k-space data, wavelet compression, and motion artifacts. Note motion artifacts were only applied to brain images. Each distortion was applied at four distinct strengths, yielding a dataset of 414 images, including the references. Three body radiologists were recruited to evaluate the quality of the abdomen stimuli, and three neuroradiologists for the brain stimuli, on a five-point diagnostic quality scale. Moreover, ten full-references objective image quality metrics were studied, i.e., root mean square error (RMSE), PSNR [22], SSIM [23], MS-SSIM [29], IW-SSIM [33], gradient magnitude similarity deviation (GMSD) [36], FSIM [30], high dynamic range visible predictor (HDR-VDP) [37], NQM [24], and VIF [25]. It is interesting to mention that no data analysis was carried out on the subjective scores; but a comparison between subjective and objective ratings was made. In general, as expected, all the image quality metrics scores improved as radiologists' scores increased. VIF, FSIM, and NQM performed statistically better than the other objective metrics (note correlation values were not given). The authors wished to highlight the fact that, even though commonly used to assess the quality of medical images, RMSE and SSIM were among the metrics with the poorest correlation with the subjective scores.

2.2. Ultrasound

In 2014, Loizou *et al.* compared eight despeckling filters applied to ultrasound images [38]. The assessment of such correlated multiplicative noise, namely speckle, is challenging since it carries useful information while altering the performance of postprocessing algorithms. Filtering indeed allows a better separation of classes between asymptomatic and symptomatic subjects. The authors recorded 100 B-mode ultrasound images of the common carotid artery, respecting Cyprus national bioethics committee rules. The B-Mode scans were acquired using Philips ATL HDI-3000 with regards to acquisition protocol, adjusted settings, and standardised post-processing method [39], for an optimal visualisation of the intima-media complex (IMC), which thickness is an early indicator of cardiovascular disease. The ultrasound scans, obtained from asymptomatic subjects, were first filtered using four categories of filtering, i.e., linear (with four filters), nonlinear (with two filters), first order statistics (with two filters), and diffusion (with two filters). Then, despeckled images were automatically segmented in order to extract the IMC thickness [40]. Multiscale texture (i.e., amplitude and frequency) was performed. The selected features were assessed for classification as regards to level of cardiovascular risk to develop. Two experts, a cardiovascular surgeon and a neurovascular specialist, were asked to assess images according to the double stimulus continuous quality scale (DSCQS) [16]. Original, and despeckled and segmented images were randomly viewed by the assessors, who had to assign a score for each image depending on their visual perception. Furthermore, a no reference image quality assessment (NR-IQA) metric, the Naturalness Image Quality Evaluator (NIQE) index [41], was used. Both subjective and objective evaluations revealed the same ranking of the despeckling filters, with the best performance for the hybrid median filter (i.e., non-linear category), and the poorest performance for first order statistics filtering. The last outcome of the paper was that the multiscale texture analysis increased the number of significant features of despeckled images and improved the class separation.

The same year, Razaak *et al.* addressed a similar issue, with an additional parameter, i.e., time dimension [42]. Indeed, their study involved the quality assessment of nine original medical ultrasound videos of heart, liver, kidney, and lung. The latter were compressed using high-efficiency video coding (HEVC), at eight different quantisation parameters (QP). This led to a total of seventy-two processed video sequences lasting four seconds each. Four medical experts, as well as sixteen viewers without medical expertise, were involved in a subjective experiment where processed videos were presented beside their original version, i.e., according to the DSCQS methodology [16]. Participants were asked to rate both sequences on two separate five-level rating scales. Prior to calculation of the DMOS, the scores of one non-expert subject were rejected as they were outside the confidence interval. Additionally, seven objective metrics were employed to compute quality, i.e., PSNR [22], SSIM [23], UQI [32], NQM [24], VIF [25], visual signal-to-noise ratio (VSNR) [43], mean square error (MSE), and video quality metric (VQM) [44]. Pearson linear correlation coefficient (PLCC),

and Spearman rank order correlation coefficient (SROCC) revealed the correlation between subjective and objective scores; VIF, UQI, and SSIM appeared to assess reliably visual and diagnostic quality (i.e., with correlation values of 0.98 for VIF, 0.96 for UQI, and 0.95 for SSIM). Moreover, analysis run on the DMOS indicated that HEVC could be used to compress ultrasound videos with low bit rate requirements without compromise diagnostic quality. Razaak *et al.* continued their work in 2016 [45]. In this latest study, they proposed a diagnostic quality-oriented video metric, namely cardiac ultrasound video quality index (CUQI). The new metric was tested on cardiac ultrasound sequences, which were also subjectively evaluated by four medical experts (i.e., three cardiologists and one radiologist), under the same conditions than in [42]. Correlation analysis showed a high correlation between CUQI values and subjective scores. Finally, CUQI was compared to the other objective metrics studied in [43]. In terms of correlation with subjective scores, CUQI was more reliable than the seven other existing metrics. In terms of significance, CUQI performance was statistically better than VIF, UQI, and PSNR; equal to VSNR, and NQM; and worse than SSIM, and VQM. The authors concluded by claiming that CUQI could be used for a reliable objective evaluation of the diagnostic quality of cardiac ultrasound videos.

Gray *et al.* explored ultrasound imaging under real time wireless transmission, with others quality issues linked to compression, bit rate, and real time [46]. For this purpose, they focused on ultrasound videos used in emergency situations for trauma identification. The video clips of six distinct anatomical areas, i.e., right and left lungs, right and left upper quadrants, heart, and pelvis were provided by the emergency department of Hackensack University Medical Centre, New Jersey, from six anonymised patients. In order to simulate video distortion caused by real time transmission, H.264 compression scheme was used considering six settings, i.e., 0.1, 0.2, 0.5, 1, 2, and 4 Mbps (highest quality, taken as reference video). The subjective evaluation was performed using the DSCQS [16] by four ultrasound trained medical professionals: one radiologist, one emergency physician, and two ultrasound engineers. The MOS revealed, based on Friedman two-way analysis of variance (ANOVA) statistical test, a significant difference among image content (i.e., anatomical area) and bit rates. Moreover, objective quality assessment was conducted with two FR-IQA metrics, namely, PSNR [22] and SSIM [23], providing numerical measures for the studied videos. All mean PSNR and SSIM values were significantly different regarding content and bit rates. The authors established a correlation between subjective and objective assessment. Finally, they set a threshold at 1 Mbps as minimum bit rate for wirelessly transmitted ultrasound videos to be of adequate quality and to allow physicians to make accurate diagnosis.

More recently, in 2018, Outtas *et al*. worked on despeckeling of liver ultrasound images [47]. The studied methods, as well as the proposed one, were subjectively and objectively tested on a parenchymal organ. Twelve liver ultrasound images were used. The subjective experiment was performed according to the subjective assessment for video quality (SAMVIQ) protocol [48]. Three radiologists with different years of experience were invited to evaluate the quality of the ultrasound images using four criteria, i.e., diagnosis, contrast, texture conspicuity, and edge sharpness. SAMVIQ interface allowed to visualise each image several times, and to re-evaluate a previously scored image, using a continuous rating scale ranging from 0 to 100, with an explicitly labelled reference. For the objective evaluation, the authors chose four NR-IQA metrics, i.e., the speckle's signal-to-noise ratio (SSNR) [49], blind image quality evaluator based on scales (BIQES) [50], NIQE [41], and NIQE-K [51]. Five filters were evaluated: the anisotropic diffusion filter with memory based on speckle statistics (ADMSS) [52], the optimised Bayesian non-local means filter with block selection (OBNLM) [53], and three outputs (i.e., texture, edge, and global enhancement) of the multi-output filter based on multiplicative multiresolution decomposition (MOF-MMD) [54]. The explicit and hidden references were used as an anchor with the sixty stimuli for the subjective assessment. MOS were calculated for each criterion across despeckling methods. Thanks to statistical analyses, the authors claimed that edge sharpness coincided with diagnostic facility, and that texture enhancement could be more or less useful depending on radiologists' experience. An ANOVA revealed no significant difference between observers in scoring image quality, while both content and despeckling method had a significant effect on the perceived quality. Finally, the correlation between subjective and objective scores was calculated using PLCC and SROCC, and revealed that NIQE was the closest metric to radiologists' opinion scores, with correlation values between 0.34 and 0.88.

Lévêque *et al.* carried out two subjective assessment studies using a same video dataset. In the first work, they aimed to investigate the impact of medical specialty settings on the perceived quality of ultrasound videos [3]; whereas, in the second work, their goal was to compare the perception of radiologists coming from two distinct continents [55]. For both studies, the authors extracted four source videos from hepatic ultrasound scans. These reference stimuli were compressed using two video codecs, i.e., H.264 and HEVC. More precisely, seven compressed sequences were generated for each reference, at the following bit rates: 512, 1000, and 1500 kbps with H.264, and 384, 512, 768, and 1000 kbps with HEVC, yielding a total of thirty-two stimuli. For the scoring interface, they adopted the SAMVIQ concept [48]; however, they created three semantic portions, i.e., "not annoying" ([75,100]), "annoying but acceptable" ([25,75]), and "not

acceptable" ([0,25]), instead of the original quality scale. For the first experiment, eight radiologists were recruited from Angers University Hospital, France, and nine sonographers (i.e., radiographers trained to perform ultrasounds) from Castle Hill Hospital and Hull Royal Infirmary, United Kingdom. An outlier detection and subject exclusion procedure was applied to the raw scores, which were then averaged towards MOS. An ANOVA revealed that the compression scheme, compression ratio, and video content affect the perceived quality for both specialty groups. As expected, perceived quality increased with the bit rate. Furthermore, videos compressed with HEVC were overall rated higher than videos compressed with H.264 at a same bit rate. In general, sonographers were more annoyed by highly distorted videos than radiologists, while no statistically significant difference was found between both groups at higher quality. For the second experiment, eight expert radiologists from Angers University Hospital, France, and five expert radiologists from Xi'an Children's Hospital were involved. Similar data analyses were carried out and showed, once again, the significant impact of compression configuration and video scene on the perceived quality. The Chinese radiologists generally gave higher scores than the French ones for all the videos. Moreover, a variation was found between both groups regarding the order they rated the video scenes, which may be due to their sensitivity to specific distortions. To conclude, with these studies, the authors' objective was to better understand the impact of medical specialty and practice settings on the perceived quality of visual content.

2.3. Computerised tomography (CT)

Liu *et al.* conducted a subjective and objective experiment to evaluate the quality of compressed computerised tomography (CT) scans [11]. They employed five neurological and five upper body slices selected from the Cancer Imaging Archive (TCIA) [56]. Each stimulus was compressed at five distinct compression ratios using two compression algorithms, i.e., JPEG and JPEG2000. Several radiologists were involved to assess the perceived quality of these distorted images (note the authors did not mention the number of participants). The pair comparison method was used; the participants were asked to give a binary answer for each stimulus, i.e., acceptable or unacceptable. Each compressed image was presented twice to the radiologists, without their knowledge. Four objective metrics were applied to the image dataset: compression ratio (CR), mean square error (MSE), quality factor (QF), and SSIM [23]. In order to compare subjective and objective scores, the authors performed a receiver operating characteristic (ROC) analysis, where they assumed the answers given by the radiologists was the ground truth. Overall, the highest areas under ROC curve (AUC), comprised between 0.93 and 0.95, corresponded to the SSIM values. Similarly, the largest Kolmogorov Smirnov (KS) values were linked to SSIM. On the contrary, CR showed the poorest performance among the four metrics studied.

Finally, Tang *et al.* conducted a relatively novel experiment with a view to evaluate the quality of multimodal medical fused images (MMIF) [57]. The fusion of medical images, e.g., MRI and positron emission tomography (PET), allows to provide complementary information about the human body. The authors constructed a new image database comprised of thirty-four pairs of original images. The latter covered different imaging modalities, i.e., MRI and computed tomography (CT), MRI-T1 and MRI-T2, ultrasound and single photon emission CT (SPECT), MRI and PET, and MRI and SPECT. A total of 272 medical fused images were generated thanks to the use of eight MMIF algorithms, i.e., image fusion with nuclear norm minimisation (NNM) [58], image fusion with multi scale transform and sparse representation (LP-SR) [59], volumetric medical image fusion with cross-scale coefficient selection (CSCS) [60], image fusion with guided filtering (GFF) [61], nonsubsampled contourlet transform (NSCT) based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency (DES) [62], medical image fusion with improved sum-modified-Laplacian (ISML) [63], image fusion with convolutional sparse representation (CSR) [64], and multimodal medical image fusion with discrete Tchebichef moments and pulse coupled neural network (DTM-PCNN) [65]. The authors recruited twenty radiologists who had to rate each medical image fused on a continuous scale between zero and five, where 0-1 meant bad image fusion performance, 1-2 poor fusion result, 2-3 fair image fusion quality, 3-4 good image fusion result, and 4-5 excellent fusion performance. MOS were generated from the obtained subjective scores. Overall, the ISML algorithm performed better (i.e., highest MOS values), followed by the LP-SR and DTM-PCNN. The CSCS algorithm showed the worst performance. In their article, Tang *et al.* also proposed a new quality metric for MMIF images. They compared it with eleven image fusion quality metrics, i.e., gradient based fusion metric [66], structured based metrics [67], edge information [68], phase congruency [69], ratio of spatial frequency error based metrics [70], mutual information based metrics [71], entropy [59], optimising structural similarity index [72], Tsallis entropy-based metrics [73], distorted image quality [74], and multi-task end-to-end optimised deep neural network [75]. The proposed scheme obtained the best performance when applied on the MMIF dataset (i.e., correlation values around 70% with MOS). However, it shall be highlighted that none of the studied state-of-the-art image fusion quality evaluation metrics was able to predict the quality of MMIF images (i.e., correlation values below 50%).

Table I: Review of the research works published in the literature on the quality assessment of radiological images and videos.

Source article	Image / video content	Distortion types and levels	Participants	Method	Scores	Objective metrics	Statistical analyses	Key findings
Chow *et al.* [2]	25 MR images of brain, abdomen, spine, and knee	6 distortion types (Rician noise, Gaussian white noise, Gaussian blur, DCT, JPEG compression, JPEG2000 compression) 5 different levels Total of 25 + 750 images = 775	28 electrical engineering research scholars	SDSCE	DMOS	SNR, PSNR, SSIM, MS-SSIM, FSIM, IFC, NQM, WSNR, VIF, VIFP, UQI, IW-PSNR, IW-SSIM	T-test, correlations, regression analysis	Rician and Gaussian white noises cause low contrast object to be less visible Gaussian blurring causes small objects and fine details to be less visible In DCT, JPEG and JPEG2000, artifacts caused by compression not clearly seen Highest correlation for NQM, lowest for UQI
Gray *et al.* [46]	Ultrasound 18-second videos of 8 trauma patient, each video composed of 6 clips (2 lungs, 2 upper quadrants, heart, pelvis)	5 H.264 bit rates (0.1, 0.2, 0.5, 1, 2 Mbps)	4 UTMPs (1 radiologist, 1 emergency physician, 2 ultrasound trained engineers)	DSCQS	Perceived quality	PSNR, SSIM	ANOVA, Tukey's multiple comparison test	Minimum bitrate threshold for wireless transmission of ultrasound video determined at 1 Mbps
Lévêque *et al.* [3]	4 hepatic ultrasound videos of 12s each, 1920×1080 pixels, 25 fps, no pathology	2 video codecs (H.264 and HEVC), 7 bit rates: 4 with HEVC (284, 512, 768, 1000kbps), 3 with H.264 (512, 1000, 1500kbps) Total of 32 stimuli	17 medical professionals: 8 radiologists, 9 sonographers	SAMVIQ, acceptability vs. annoyance	MOS	/	ANOVA Post-hoc tests	Significant impact of visual content and compression configuration on perceived quality Video quality changes with content and compression configuration tend to be consistent for radiologists and sonographers
Lévêque *et al.* [55]	4 hepatic ultrasound videos of 12s each, 1920×1080 pixels, 25 fps, no pathology	2 video codecs (H.264 and HEVC), 7 bit rates: 4 with HEVC (284, 512, 768, 1000kbps), 3 with H.264 (512, 1000, 1500kbps) Total of 32 stimuli	8 expert radiologists from France, 5 expert radiologists from China	SAMVIQ, acceptability vs. annoyance	MOS	/	ANOVA Post-hoc tests	Video scene and compression configuration (codec and bit rate) affect perceived quality Quality increases with bit rate, HEVC better than H.264 Strong agreement between radiologists within each group, differences between groups (Chinese scored higher)
Liu *et al.* [34]	3 MR images: 2 brains, 1 liver	4 types of artifacts: white noise, coloured noise, edge ghosting, ghosting 5 levels of energy Total of 60 stimuli	15 clinical scientists and applications specialists	Simultaneous double stimulus (SDS)	MOS linearly remapped to [1, 10]	/	ANOVA Post-hoc tests	Image content, artifact type, and energy level affect quality Small difference in perceived quality between ghosting and white noise (ghosting scored higher) For liver, white noise scored lower than ghosting Coloured noise consistently scored higher than edge ghosting
	8 MR images (2 brains, 1 liver, 1 breast, 1 foetus, 1 hip, 1 knee, 1 spine)	4 different versions of coloured noise 2 levels of energy Total of 112 stimuli	18 clinical scientists and applications specialists	Simultaneous double stimulus (SDS)	MOS linearly remapped to [1, 10]	/	ANOVA Post-hoc tests	

Liu *et al.* [11]	5 neurological and 5 upper body CT slices	2 compression types: JPEG and JPEG2000 5 compression ratios	Radiologists	Binary: acceptable, unacceptable	Perceived quality	CR, MSE, QF, SSIM	ROC AUC, KS	Both using AUC and KS, SSIM performed better CR demonstrated worst performance QF provided moderately reasonable predictions on JPEG MSE performed inconsistently
Loizou *et al.* [38]	100 B-mode longitudinal ultrasound images of common carotid artery	AM-FM simulation	2 medical experts: 1 cardio-vascular surgeon, 1 neuro-vascular specialist	DSCQS	Perceived quality	NIQE	Mann-Whitney rank sum test, ranking	Filtering improve class separation between asymptomatic and symptomatic subjects NIQE ranking similar to visual experts ranking
Mason *et al.* [35]	9 MR images of brain and 9 of abdomen, no pathology GE 3T MR 750 Discovery or 1,5T Signa HDxt scanner	White Gaussian noise, Gaussian blurring, Rician noise, under sampling, motion artifacts, wavelet compression Applied with varying strengths (motion artifacts only to brain images) Total of 414 stimuli	3 body radiologists (abdomen images), 2 neuro radiologists (brain images)	1-5 Likert scale	Diagnostic quality	RMSE, SSIM, PSNR, MS-SSIM, IW-SSIM, GMSD, FSIM, HDRVP, NQM, VIF	Raw scores converted to difference scores then to z-scores SROCC Variance-based hypothesis test	As radiologists' scores increase, all IQ metrics scores tend to improve When dividing data by radiologist and by reference image; VIF, FSIM, and NQM perform better When dividing by degradation, variation is less clear SSIM and RMSE do not show strong correlation with experts
Outtas *et al.* [47]	21 in vivo abdominal liver (granular, smooth, cirrhotic, non-cirrhotic) ultrasound images	Speckle noise	3 radiologists with different years of experience	SAMVIQ 4 criteria: contrast, ability to diagnose, texture conspicuity, edge sharpness	MOS	SSNR, NIQE, NIQE-K (proposed metric), BIQES	ANOVA SROCC, PLCC	Proposed filter allowed image enhancement
Rajagopal *et al.* [19]	10 MR images	4 distortion levels: Rician noise, Gaussian white noise, Gaussian blur, DCT compression Total of 210 images	10 electrical engineering research scholars	SDSCE	DMOS	PSNR, SSIM, NQM, VIF	Descriptive statistics, correlations, regression analysis	Subjective quality scores close to objective metrics
Razaak *et al.* [42]	9 ultrasound videos (3 hearts, 3 livers, 2 kidneys, 1 lung), 25 fps	8 HEVC quantisation parameters (27, 29, 31, 33, 35, 37, 39, 41) Total of 92 videos	4 medical experts, 16 naïve participants	DSCQS Diagnostic quality for specialists, perceived quality for non-experts	DMOS	PSNR, SSIM, UQI, VQM, NQM, VIF, VSNR	Correlations, regression analysis	Highest correlations with DMOS of experts for VIF, UQI, and SSIM Diagnostically reliable videos can be obtained for compression up to QP=35 with HEVC

Razaak *et al.* [45]	3 cardiac ultrasound videos, 640×416, 25 fps	8 HEVC quantisation parameters (27, 29, 31, 33, 35, 37, 39) Total of 27 videos	4 medical experts (3 cardiologists, 1 radiologist)	DSCQS type II	DMOS	CUQI (developed by the authors), SSIM, VSNR, VIF, UQI, PSNR, VQM, NQM	Correlations, regression analysis	CUQI performed better than some state-of-the-art metrics SSIM, VQM, VIF showed high performance
Suad *et al.* [17]	20 MR images, 512×512	5 distortion types: Gaussian noise, blurring, JPEG compression, salt and pepper, sharpness Total of 100 images	15 specialists in electronic medical diagnosis	DSIS	MOS	/	Descriptive statistics	Sharpness shows highest quality ratio, Gaussian noise poorest quality
Tang *et al.* [57]	34 pairs of fused images: MRI and CT, MRI-T1 and MRI-T2, ultrasound and SPECT, MRI and PET, MRI and SPECT	8 MMIF algorithms: NNM, LP-SR, CSCS, GFF, DES, ISML, CSR, DTM-PCNN	20 radiologists	0-5 continuous quality scale	MOS	Proposed metric, 11 state-of-the-art image fusion quality metrics	Correlations	ISML algorithm led to highest MOS values, CSCS led to lowest values Metric proposed by the authors showed higher correlation with MOS than existing metrics

Legend: 3D (three dimensional); AM-FM (Amplitude-Modulation Frequency-Modulation); ANOVA (Analysis of Variance); AUC (Area Under Receiver Operating Characteristic Curve); BIQES (Blind Image Quality Evaluator based on Scales); CNN (Convolutional Neural Network); CR (Compression Ratio); CSCS (Cross-Scale Coefficient Selection); CSR (Convolutional Sparse Representation); CT (Computed Tomography); CUQI (Cardiac Ultrasound Video Quality Index); DCT (Discrete Cosine Transform); DES (Pulse-Coupled Neural Network and Modified Spatial Frequency); DL (Deep Learning); DMOS (Differential Mean Opinion Score); DSCQS (Double Stimulus Continuous Quality Scale); DSIS (Double Stimulus Impairment Scale); DTM-PCNN (Discrete Tchebichef Moments and Pulse Coupled Neural Network); FSIM (Feature Similarity Index Measure); GFF (Guided Filtering); GMSD (Gradient Magnitude Similarity Deviation); HDR-VDP (High Dynamic Range Visible Difference Predictor); HEVC (High Efficiency Video Coding); IFC (Information Fidelity Criterion); ISML (Improved Sum-Modified-Laplacian); IW-PSNR (Information Weighted Content Peak Signal-to-Noise Ratio); IW-SSIM (Information Weighted Content Structural Similarity Index Measure); KS (Kolmogorov-Smirnov test); LP-SR (Multi Scale Transform and Sparse Representation); MMIF (Multimodal Medical Fused Images); MOS (Mean Opinion Score); MR (Magnetic Resonance); MSE (Mean Square Error); MS-SSIM (Multi-Scale Structural Similarity Index Measure); NIQE (Natural Image Quality Evaluator); NNM (Nuclear Norm Minimisation); NQM (Noise Quality Measure); PET (Positron Emission Tomography); PLCC (Pearson Linear Correlation Coefficient); PSNR (Peak Signal-to-Noise Ratio); QF (Quality Factor); QP (Quantisation Parameter); RMSE (Root Mean Square Error); ROC (Receiver Operating Characteristic); SAMVIQ (Subjective Assessment Methodology for Video Quality); SDSCE (Simultaneous Double Stimulus for Continuous Evaluation); SPECT (Single Photon Emission Computed Tomography); SSIM (Structural Similarity Index Measure); SNR (Signal-to-Noise Ratio); SROCC (Spearman's Rank Correlation Coefficient); UTMP (Ultrasound Trained Medical Professionals); UQI (Universal Image Quality Index); VIF (Visual Information Fidelity); VIFP (Visual Information Fidelity in Pixel domain); VSNR (Visible Signal-to-Noise Ratio); WSNR (Weighted Signal-to-Noise Ratio).

3. Review of the perceptual-based approaches used for the assessment of medical image quality in surgery and other practices

This section concerns articles on medical image quality assessment in surgery and other practises, which are summarised in Table II.

3.1. Endoscopic, laparoscopic, and open surgeries

A decade ago, Nouri *et al*. published a short paper in which they aimed to define compression thresholds for the telesurgery context [76]. To this purpose, they extracted four videos from different operations, including a movement of surgical laparoscopic instrument, a blood clot, a fat area, and a compress application. Each of these sequences were compressed using MPEG-2 standard at several ratios (i.e., bit rates ranging from 1.02 to 7.2 Mbps). A total of twenty-five distorted videos were generated. The DSCQS [16] method was chosen for the subjective quality evaluation, carried out with seven expert surgeons. Note one surgeon was removed from the database as their scores were considered atypical. Two video sequences (i.e., blood clot and fat area), were not strongly affected by the degradation, even at the lowest bit rate studied (i.e., 1.5 Mbps). The other two sequences (i.e., surgical instrument and compress application) were, on the contrary, impacted by bit rate variation. The authors performed a regression analysis and observed that, above 3.2 Mbps, no loss of quality was perceived by the surgeons. They concluded by determining a compression threshold at 90:1 for the field of robotic-assisted surgery.

Münzer *et al.* also investigated the impact of compression on laparoscopic videos, but this time with a view to define archiving recommendations [77]. They collected forty-eight distinct full HD laparoscopic videos, representing various procedures and phases (e.g., overview, cutting, suturing). The H.264 coding standard was used to compress the original video sequences at several constant rate factors (CRF), comprised between 18 and 28 (note a higher CRF value is linked to a lower quality). Two test sessions were organised by the research team, who recruited eighteen surgical residents as well as nineteen experienced surgeons. The first session, which aimed to evaluate to what extent could laparoscopic videos be compressed without perceived quality loss, was conducted using the DSCQS [16]; while the second session, intending to study the impact of technical on semantic quality, was performed thanks to the absolute category rating hidden reference (ACR-HR) method [15]. Twelve distinct test conditions, corresponding to different CRF and different resolutions (i.e., 1920×1080, 1280×720, 960×540, and 640×360), were evaluated within each session. Raw data collected was pre-processed using the outlier detection method, yet no outliers were identified. DMOS were calculated from the scores obtained during the first session. It is interesting to note that, with full HD resolution (i.e., 1920×1080) and for a CRF between 20 and 26, distorted videos were rated higher than references (i.e., negative DMOS). A significant increase of the DMOS was observed for the three lower resolutions, corresponding to a clear quality deterioration. Data obtained during the second session led to MOS calculation. Satisfactory scores were obtained under all tested conditions, meaning low technical quality provided acceptable semantic quality. However, decrease of resolution yielded lower MOS, even for higher bit rates. Finally, Münzer *et al*. investigated the impact of experience on perceived quality. They showed that, in general, experienced surgeons had lower requirements than residents. The authors concluded their work by providing guidelines on H.264 compression for laparoscopic video.

Two articles were published by Chaabouni *et al*. with a yearly interval to describe their own study [78]-[79]. They acquired four original video sequences from an ear, nose, and throat (ENT) surgery performed at Nancy University Hospital, France. The four excerpts were encoded with H.264 standard at eleven compression ratios. Fourteen medical professionals, having different ENT experiences (i.e., intern, extern, resident, doctor, professor), were involved in a subjective experiment, carried out using the DSCQS method [16]. The MOS were generated for each video sequence. Using a regression analysis, the authors defined a compression bit rate threshold around 10 Mbps with H.264 for the first sequence. An objective evaluation was conducted on top of the subjective one, where the scores of thirteen metrics, i.e., PSNR [22], MSE, SSIM [23], MS-SSIM [29], UQI [32], IFC [31], VIF [25], VIFP [25], PSNR human visual system (PSNR-HVS) [61], PSNR-HVS-M [80], HDR-VDP [37], blind referenceless image spatial quality elevator (BRISQUE) [81], and NIQE [41] were investigated. The correlations between subjective (i.e., MOS) and objective measures for each stimulus were calculated using Pearson and Spearman coefficients. Results revealed MSE (correlation between 0.93 and 0.99), NIQE (between 0.92 and 0.98), NQM (between 0.91 and 0.98), SSIM (between 0.89 and 0.96), MS-SSIM (between 0.91 and 0.94), and BRISQUE (between 0.92 and 0.97) as best metrics. Furthermore, Chaabouni *et al.* performed a preliminary study using HEVC standard. They compared the values obtained with four objective metrics, i.e., PSNR, MSE, MM-SSIM, and NQM, on videos encoded with H.264 and HEVC. They concluded that, as expected, HEVC seemed to give better results than its predecessor.

Kumcu *et al.* conducted their own study to assess the quality of full HD compressed laparoscopic videos [82]. They captured four abdominal test sequences using two distinct camera systems, i.e., a standard optical one and a chip-on-tip digital camera. The four videos were compressed with H.264, at four different bit rates (i.e., 1.8, 2.8, 5.5, and 19.5 Mbps, respectively corresponding to compression ratios 336, 214,

111, and 31). Therefore, twenty stimuli were generated in total, including the references. Nine expert laparoscopic surgeons from Ghent University Hospital, Belgium, and sixteen doctoral and post-doctoral researchers were recruited for a video quality assessment study using the single stimulus continuous quality evaluation (SSCQE) [16]. Both groups were asked to rate the quality of each sequence, and the medical experts were also invited to evaluate the "suitability for surgery". An outlier detection was conducted, and the remaining raw scores were converted to subjective difference median opinion scores (DMdnOS). First statistical analyses showed that quality scores given by surgeons were inversely correlated to compression bit rates. Furthermore, their scores were significantly impacted by the scene. It is interesting to note that the ratings obtained for "suitability for surgery" yielded more significant differences than those for quality, and that the 19.5 Mbps compressed sequence was rated as more suitable than the reference by the surgeons. On the contrary, the non-expert group rated differently the videos. The authors concluded that a bit rate of 5.5 Mbps could be suitable for surgical procedures, and that non-experts should not be involved to gauge surgeons' preferences. In addition, three objective quality assessment metrics were evaluated, i.e., VQM [44], PSNR [22], and HDR-VDP-2 [37]. VQM was the only metric which demonstrated an acceptable correlation with the subjective scores. Three years later, the research team re-used their dataset comprised of laparoscopic video sequences with a view to evaluate four subjective quality assessment protocols [83]. More precisely, they aimed to compare the performance of forced choice preference, ratio-scaled preference, quality dissimilarity, and single stimulus quality protocols, as well as of diverse rating pre-processing approaches, resulting to a total of fourteen distinct methods. To do so, they recruited ten laparoscopic surgeons from Ghent University Hospital, and seventeen doctoral and post-doctoral researchers. Overall, the ranking of the mean quality scores was consistent across the studied analysis methods. The preference and dissimilarity protocols, associated with multidimensional scaling (MDS) analysis, and the single stimulus protocol, associated with z-score analysis, performed better than the other methods. Kumcu *et al*. noted that the dissimilarity protocol with MDS only required a few participants to achieve enough study power.

In 2017, Lévêque *et al*. investigated video quality in the particular context of telesurgery [8]. Thanks to semi-structured interviews carried out with expert abdominal surgeons, they designed and conducted a controlled subjective experiment. Two distinct surgical procedures were studied, namely, open and laparoscopic surgeries. For each procedure, four videos were extracted from different surgical acts. The eight sequences were compressed with H.264, at five bit rates (i.e., 128, 256, 350, 512, and 1000 kbps). Furthermore, packet losses were generated to simulate transmission errors, at two distinct rates (i.e., 1, and 3%). Finally, original frame rate was divided by two (i.e., from 30 to 15 fps for open surgery, and from 25 to 12.5 fps for laparoscopy). The video dataset was consequently comprised of a total of thirty-two distorted videos (i.e., eight distorted conditions per reference) for each procedure. A total of eight surgeons, i.e., four experts in open surgery and four in laparoscopy, from Angers University Hospital, France, participated in the subjective experiment. They were asked to rate the perceived quality of all distorted videos, in the absence of references, through a single-stimulus method [16]. More specifically, they were invited to consider the quality in terms of suitability to help a remote surgeon. Additionally, they were asked to rate four other criteria, i.e., colours, contrasts, reliefs, and textures. Outlier detection and subject exclusion procedures were applied to the raw data for both surgical procedures. Correlation analyses, using the PLCC, were then conducted to quantify the links between overall quality and each studied criterion. As the correlation values were very high (i.e., between 89 and 97%), only the quality scores were considered and further converted to MOS. For both procedures, an ANOVA showed the significant effect of video content, bit rate, and packet loss rate on the perceived quality. The impact of the frame rate was not significant in the case of open surgery, while it was significant for laparoscopy. Lévêque *et al*. therefore claimed that perceived quality was dependent on the specific procedure studied. The same year, the authors also published another research work using the same four original video sequences of open surgery [84]. However, they used different distortion types, i.e., they compressed the excerpts with H.264 and HEVC. For each source, seven compressed versions were generated, i.e., 256, 384, and 512 kbps using H.264, and 128, 256, 384 and 512 kbps using HEVC, which led to a total of thirty-two video stimuli. The authors employed the same scoring methodology than in [3], i.e., an adapted version of SAMVIQ [48]. Eight abdominal surgeons from Angers University Hospital were recruited for this experiment. The study of the MOS illustrated the impact of the compression scheme, bit rate, and content on the perceived quality. The authors concluded that, at low quality, the bit rate of H.264 should be 2 times as high as that of HEVC to obtain a similar perceived quality; while, at higher quality, the bit rate of H.264 should be 1.5 times that of HEVC.

The same year, Usman *et al*. investigated the context of wireless video transmission [6]. They were then the first team that studied video quality assessment for wireless capsule endoscopy (WCE), a swallowable capsule that contains a camera and a small circuit. Such captured videos are usually transmitted and analysed for the diagnosis of gastro-intestinal abnormalities. The authors performed an extensive subjective and objective quality evaluation to investigate the suitability of HEVC standard compression. The perceptual experimentation involved six experts and nineteen non-

experts, consisting in research students in wireless emerging networks. To make the experiment more substantial, ten video sequences were chosen according to their content complexity (i.e., spatial and temporal information). Impaired videos were obtained through eight different compression levels with HEVC encoder (i.e., QP = 27, 29, 31, 33, 35, 37, 39, 41). Visual quality evaluation was conducted using DSCQS type-II [15], with a software developed by [85]. Before conducting the subjective tests, a training session allowed the authors to guarantee the inter-rater reliability of the results. Hence, three tasks were designed, aiming to observe scoring consistency evaluating different levels of compression of sequences belonging to the same disease, identify participants who scored the clips based on content by evaluating ten distinct compressed videos at a same level (i.e., QP=35), and see whether observers differentiate compression levels by evaluating five videos, each compressed at two levels. No outliers were detected for the expert group, whereas one outlier was removed among the non-expert group. Ten VQA metrics were further tested on the video dataset, i.e., PSNR [22], SSIM [23], MS-SSIM [29], VSNR [43], IFC [31], VIF [25], VIFP [25], UQI [32], NQM [24], and WSNR [24]. The DMOS analysis revealed an upper limit for HEVC compression, i.e., QP = 37, for WCE videos considering different gastro-intestinal diseases. Furthermore, the authors concluded that non-expert observers presented a large variation, and could only be considered for visual quality evaluation; expert opinion is the most suitable for the assessment of medical video content. VIF, VIFP, and IFC showed the best correlation with subjective measurements, (i.e., between 88 and 92%), with an outperformance of VIFP in terms of statistical significance and computation time.

Khan *et al.* recently released a publicly available database on the quality assessment of laparoscopic videos [86]. They proposed a computational framework that combines two modules, i.e., one for video quality assessment, and one for enhancement. Their database consists of ten reference videos of ten seconds each. All the videos were distorted using five distortion types: smoke, noise, uneven illumination, blur due to defocus, and blur due to motion, each at four different levels. A subjective experiment was conducted following the pairwise-comparison (PC) protocol, with two separate groups of observers. Indeed, thirty non-experts and ten experts were invited to randomly evaluate all possible pair combinations of videos from a same category. MOS analysis showed that experts differently (i.e., significantly worse) perceived the quality than non-experts for all distortion types, except for defocus blur. However, more pronounced was the distortion, less was the perceptual difference between experts and non-experts. This can be explained by the fact that experts are more task-oriented. An objective assessment was also performed with three FR metrics, i.e., PSNR [22], SSIM [23], and VIF [25], as well as three NR metrics, i.e., BRISQUE [81], NIQE [41], and VIIDEO [87]. Both PLCC and SROCC showed that VIF had a maximum correlation with non-experts, whereas PSNR was more correlated with experts. However, when distortions were individually considered, VIF correlated much better with the subjective scores of all assessors. The authors finally specified that both NIQE and BRISQUE are adapted for noise and defocus blur estimations.

3.2. Other medical visual content

In 2017, Kara *et al.* addressed the issue of three-dimension (3D) medical image quality using autostereoscopic display principle (i.e., without 3D glasses) [88]. In this context, it is important to ensure a continuous motion parallax helping a sufficient visual input to avoid image alteration and flawed diagnosis. The authors studied light field reconstruction of intermediate views. To do so, they conducted two experiments using LED-based 3D projection unit with a forty-degree field of view of human heart. The experiments involved twenty participants, including eight experts and twelve non-experts. For both experiments, they were asked to slightly move to the left- and right-hand sides to properly observe the stimuli, which they had to score thanks to a discrete ACR scale [15]. The first experiment aimed to evaluate angular resolution; ten views were selected, running from views 15 to 150. For the second experiment, six stimuli were assessed: three were selected from the first experiment, and three were created with light field reconstruction using Shearlet transform, with a given decimation factor. From the recorded MOS, the authors found sixty views as the minimum angular resolution for an acceptable quality. The other outcome of the study was that a higher factor decimation could provide a better visual quality. The authors concluded that a lower number of views less alters quality than degradations in texture due to view synthesis.

Table II: Review of the research works published in the literature on the quality assessment of surgical and other medical images and videos.

Source article	Image / video content	Distortion types and levels	Participants	Method	Scores	Objective metrics	Statistical analyses	Key findings
Chaabouni *et al.* [78]-[79]	4 10-second endoscopic videos from ENT surgery	11 H.264 compression ratios Preliminary study with HEVC (only with objective metrics)	14 observers with different years of experiences in ENT (interns, externs, residents, doctors, professors)	DSCQS	MOS	H.264: SSIM, UQI, PSNR, WSNR, VSNR, HDR-VDP, IFC, MSE, MS-SSIM, NIQE, NQM, PSNR-HVS, PNSR-HVS-M, VIF, VIFP, BRISQUE HEVC: PSNR, MSE, MMSIM, NQM	Correlation, regression analysis	Video could be lossy encoded from 100:1 up to 270:1, maintaining observer satisfactions Best metrics: MSE, NIQE, NQM, SSIM, MS-SSIM, BRISQUE HEVC gives better results than H.264
Kara *et al.* [88]	10 3D still images of rendered human heart	Angular resolution, light field reconstruction	8 medical experts, 12 naive observers	ACR	Perceived quality	/	Regression analysis	Observers more sensitive to degradations in texture than to lower number of views
Khan *et al.* [86]	10 10-second original laparoscopic cholecystectomy videos, each, 512×288, 25 fps	5 distortion types: smoke, noise, uneven illumination, blur due to defocus, blur due to motion, 4 different levels Total of 200 videos	10 experts, 30 naïve observers	Pairwise comparison (simultaneous presentation)	MOS	Detection: PBI, SAN, fast noise variance estimator, LMR Evaluation: PSNR, SSIM, VIF, VIIDEO, BRISQUE, NIQE	Outliers detection PLCC, SROCC	Experts perceive quality differently for all distortions compared to non-experts Even slightest level of distortion affects experts None of the metrics performed well
Kumcu *et al.* [82]	4 laparoscopic surgery videos, 1920×1080	4 H.264 bit rates (1.8, 2.8, 5.5, 19.5 Mbps)	9 laparoscopic surgeons, 16 naïve observers	SSCQE: quality and suitability	DMOS	VQM, HDR-VDP-2, PSNR	Friedman test, correlations, Wilcoxon rank-sum test, regression analysis	Video may be lossy compressed up to 100 times without sacrificing quality Surgeons sensitive to content but variance in quality score Non-experts non-sensitive to content
Kumcu *et al.* [83]	4 10-second laparoscopic surgery videos, 1920×1080	4 H.264 bit rates (1.85, 2.9, 5.6, 20 Mbps) Total of 20 stimuli	10 medical imaging experts, 17 naïve observers	Forced choice, preference, dissimilarity, SS	INDSCAL, OS, DOS, z-scores, adjacent scores, raw OS	/	BT, Wilcoxon, LME, MDS	Wide range of performance across subjective QA methods as well as within a method Ratio-scaled paired comparison methods suit small differences in quality levels
Lévêque *et al.* [8]	4 open surgery videos at 30 fps, 4 laparoscopic surgery videos at 25 fps	4 H.264 bit rates (128, 256, 350, 512, 1000kbps), 2 frame rates for open surgery (15, 30 fps), 2 frame rates for laparoscopic surgery (12,5, 25 fps) 3 packet loss rates (0, 1, 3%) Total of 64 videos	4 expert surgeons for each procedure	SS Colours, contrast, relief, texture, overall quality	MOS	/	Correlations, ANOVA	For both procedures, significant effect of compression on perceived quality For open surgery, the way the video quality changes with the bit rate depends on video content For laparoscopic surgery, impact of different bit rates on video quality is the same for all scenes

Lévêque *et al.* [84]	4 open surgery videos, 30 fps	2 video codecs (H.264 and HEVC), 7 bit rates: 4 with HEVC (128, 25, 384, 512kbps), 3 with H.264 (256, 384, 512 kbps) Total of 32 stimuli	8 abdominal surgeons	SAMVIQ, acceptability vs. annoyance	MOS	/	ANOVA Post-hoc tests	Significant impact of visual content and compression configuration on perceived quality Quality increases with bit rate, HEVC better than H.264
Münzer *et al.* [77]	48 10-second laparoscopic videos, 25 fps	6 H.264 CRF (18, 20, 22, 24, 26, 28), 4 resolutions (1920×1080, 1280×720, 960×540, 640×360)	37 medical experts (19 experienced surgeons, 18 surgical residents)	DSCQS	DMOS	/	Descriptive statistics	Possible to compress laparoscopic videos without compromising perceived quality
				ACR-HR	MOS	/	Descriptive statistics	Low technical quality still provides acceptable semantic quality, sufficient for archiving
Nouri *et al.* [76]	4 telesurgery video sequences	7 MPEG2 bit rates (from 1 to 7.2 Mbps)	7 expert surgeons	DSCQS	MOS	/	Regression analysis	A threshold above which no surgeon perceived any loss of quality was determined around 3 Mbps
Usman *et al.* [6]	10 10-second endoscopic videos, 3 fps	8 HEVC quantisation parameters (27, 29, 31, 33, 35, 37, 39, 41)	6 experienced medical doctors, 19 naïve observers	DSCQS type-II	DMOS	MSE, PSNR, SSIM, MS-SSIM, VSNR, IFC, VIF, VIFP, UQI, NQM, WSNR	Outliers detection, correlations	Videos compressed within QP range 27–31 exhibit same visual quality (no loss) MS-SSIM, VIF, VIFP, UQI, IFC showed better performance

Legend: 3D (three dimensional); ACR (Absolute Category Rating); ACR-HR (Absolute Category Rating Hidden Reference); ANOVA (Analysis of Variance); BRISQUE (Blind Referenceless Image Spatial Quality Evaluator); BT (Bradley Terry); CRF (Constant Rate Factor); DMOS (Differential Mean Opinion Score); DOS (Difference Opinion Score); DSCQS (Double Stimulus Continuous Quality Scale); ENT (Ears, Noise, Throat); HDR-VDP (High Dynamic Range Visible Difference Predictor); HEVC (High Efficiency Video Coding); IFC (Information Fidelity Criterion); IQR (Median and Interquartile Range); LME (Linear Mixed-Effect); LMR (Luminance Mean to Range); MDS (Multidimensional Scaling); MOS (Mean Opinion Score); MPEG2 (Moving Picture Experts Group 2); INDSCAL (Individual Difference Scaling); MSE (Mean Square Error); MS-SSIM (Multi-Scale Structural Similarity Index Measure); NIQE (Natural Image Quality Evaluator); NQM (Noise Quality Measure); OS (Opinion Score); PBI (Perceptual Blur Index); PLCC (Pearson Linear Correlation Coefficient); PSNR (Peak Signal-to-Noise Ratio); PSNR-HVS (Peak Signal-to-Noise Ratio Human Visual System); QA (Quality Assessment); QP (Quantisation Parameter); SAMVIQ (Subjective Assessment Methodology for Video Quality); SAN (Saturation Analysis); SS (Single Stimulus); SSCQE (Single Stimulus Continuous Quality Evaluation); SSIM (Structural Similarity Index Measure); SROCC (Spearman's Rank Correlation Coefficient); UQI (Universal Image Quality Index); VIF (Visual Information Fidelity); VIFP (Visual Information Fidelity in Pixel domain); VQM (Video Quality Measurement); VSNR (Visible Signal-to-Noise Ratio); WSNR (Weighted Signal-to-Noise Ratio).

4. Review of task-based approaches used for the assessment of medical image quality

Research works using task-based approaches for medical image quality evaluation are presented in this section.

4.1. Introduction of task performance

No recommendation has ever been given on medical subjective image and video quality assessment, although several recommendations have already been made for natural visual content [14]-[16]. This can explain why most existing subjective studies, as the ones mentioned above, still use the methodologies proposed for natural images and videos. It can be noticed that both single-stimulus (e.g., ACR [15]), and multi-stimulus (e.g., DSCQS [16]) methodologies were used, and this for different acquisition modalities (e.g., ultrasound, and endoscopy). Advantages and drawbacks exist for each methodology. For instance, single stimulus methods allow a quicker evaluation and avoid potential vote inversions when compared to double-stimulus methods [83], yet they may lead to score drift during an experiment [89]. The SAMVIQ methodology [48] tends to combine the advantages of aforementioned approaches, and can require up to thirty percent fewer observers than the ACR method [90].

A question can be raised: are these methodologies adapted for medical images and videos? It lacks studies exploring this issue. Different from natural content, that are often used for the pleasure of end-users, medical content is indeed generally used by medical experts for a specific task (e.g., a diagnostic task (detection, localisation, characterisation), or a surgical task (planning, guiding, intervention)). Therefore, a task-based subjective test methodology might be more adapted for medical image and video quality assessment. The underlying paradigm is to quantify the quality of a particular visual content by its effectiveness with respect to its intended task [91]. In addition, according to medical experts', it appears easier for them to perform the intended task on medical images (i.e., as they do in daily routine), than to judge their quality.

For subjective experiments conducted under a task-based configuration, as presented in the next subsections, medical experts are usually asked to perform one or several tasks given different medical imaging systems, and systems allowing medical experts to obtain the best task performance are said to be better. Significant work has been carried out to quantify human observer performance. In [92]-[93], the receiver operating characteristic (ROC) analysis is described, and examples of studies in radiology are reviewed.

4.2. Detection task

In 2013, Kalayeh *et al.* worked on the channelised Hotelling observer (CHO) approximation to predict human performance in a cardiac perfusion-defect detection task on single photon emission computed tomography (SPECT) images [94]. Two supervised learning regression models were considered, i.e., the channelised relevance vector machine (CRVM), and the multi-kernel channelised relevance vector machine (MKCRVM). Both methods were compared to the traditional CHO, and to a previously proposed channelised support vector machine (CSVM). As a result, MKCRVM showed the best performance in terms of accuracy, computational complexity, and execution time. Regarding the area under ROC curve (AUC), MKCRVM outperformed both CRVM and CSVM methods. Eventually, the authors noticed that all considered learning methods outperformed the classical CHO.

As several studies dealt with learning-based model observers (e.g., Kalayeh *et al.* [94]), Lorente *et al*. addressed in 2014 an important issue related to the selection of the dataset used to train a learning model observer [95]. Actually, they proposed an approach based on active learning to select data to be evaluated by human observers, and then used to train the model observer. They conducted experiments with six human observers evaluating perfusion defect visibility on simulated SPECT myocardial perfusion acquisition with eighteen reconstruction strategies. The effectiveness of the algorithm was then evaluated using the AUC, which showed an excellent prediction of human observer performance.

The same year, Marin *et al.* proposed and assessed two model observers for the prediction of human observer performance in detecting cardiac-motion defects on SPECT images [96]. The first model uses a Hotelling linear discriminant and features based on cardiac motion; while the second is based on relevance vector machines (RVM) for regression, using features from image intensity and estimated cardiac motion. To obtain the simulated data, the authors used a mathematical cardiac torso (MCAT) phantom [97], and added acquisition noise to the images. Three reconstruction methods were used, based on a filtered back-projection. Five readers participated in an observer study where they were asked to rate 180 images per reconstruction method. Overall, the RVM model showed a good correlation with human observer performance, whereas the Hotelling observer revealed a poor match.

In 2018, Wen *et al.* extended their 2D multi-lesion CHO into a 3D multi-lesion CHO [98]. Based on implementations of 3D partial least squares (PLS) and modified Laguerre-Gauss (LG) channels, this new model observer aimed to detect multiple lesion from volumetric digital breast tomosynthesis (DBT). With a view to develop such a model, they scanned breast phantoms by simulating DBT scanners of four distinct geometries. A total of 5000 lesion-free phantoms were generated. Synthetic breast lesions were integrated within four different breast areas. The model observer had to perform a multi-lesion detection task by making an image-level decision (i.e., "lesion-present" or "lesion-absent"), and a location-specific decision. To measure the observer performance, two figures of merit (FOMs) were used: the task signal-to-noise

ratio (SNR) [22], and the area under alternative free-response receiver operating characteristics (AFROC) curve [99]. Results showed that a good detection could be achieved with the 3D multi-lesion CHO with a small number of channels, and that 3D PLS channels performed better than LG channels.

More recently, Zhou *et al*. proposed to approximate the ideal observer (IO) and the Hotelling observer (HO) for the binary signal detection tasks by using of supervised learning methods [100]. To this aim, they employed artificial neural networks, i.e., for approximating the IO and HO by convolutional neural network (CNN) for the IO, and single layer neural network (SLNN) for the HO. Performances of the developed observers were evaluated using the ROC method, and compared to the performance of traditional observers or analytical calculations when feasible, i.e., without any human observer evaluation.

In the same scope of learning numerical observers, He *et al*. [101] studied the deep learning method training's issue. In fact, the availability of a large amount of labelled experimental data is not always guaranteed. The authors proposed to train a numerical observer on computer-simulated images, and to operate on experimental ones using adversarial domain adaptation methods. They provided a proof-of-principle by employing a CNN observer to learn and to perform the signal detection task with enough confidence.

4.3. Detection and localisation task

For the first time, in 2012, Zhang *et al.* proposed a numerical observer for the detection-localisation of multiple signals; the perceptually relevant channelized joint observer (PCJO) for the detection of multiple sclerosis (MS) lesions on magnetic resonance (MR) images [102]. These tasks are achieved following two main steps: a global search to locate the abnormality candidates for the localisation task, and an interpretation and cognitive analysis of each candidate to perform the detection task. The authors introduced the perceptual difference map in the computation of their numerical observer. Six radiologists with different years of experience, including two MS experts with respectively twenty-one and ten years of experience, were involved to subjectively assess ninety images on which MS lesions were simulated. The Jackknife free-response receiver operating characteristic (JAFROC) FOM indicated that the PCJO was close to radiologists' performance for the detection localisation task. Xu *et al*. extended the initial PCJO, and used it for the detection-localisation task on low dose CT images [103]. They compared the performance of two reconstruction algorithm, i.e., filtered back projection (FBP) and adaptive statistical iterative reconstruction (ASiR), with four radiologists' performance. The authors concluded that ASiR yielded better image reconstruction, and that there was no significant difference between the PCJO and the performance of the radiologists (including two CT experts).

In 2013, Leng *et al.* investigated the tasks of lesion detection and localisation on computed tomography (CT) imaging [104]. More precisely, they examined the correlation between model and human observer performance. In order to do so, the authors first scanned a water phantom containing rods to simulate low-contrast lesions of different sizes. A total of eight studies were conducted, with four dose levels and two lesion sizes. Three expert medical physicists were recruited to perform the lesion detection and localisation task on 150 images (100 with lesion, 50 without), and to rate their confidence on a 6-point scale. A CHO model observer with Gabor channels also analysed the same images (note internal noise was generated for the model observer study). Both ROC and localisation receiver observer performance (LROC) analyses were carried out for the human and model observer studies. Overall, AUC values obtained with CHO with Gabor channels were well correlated with those of human observers, which demonstrates the ability of such model to assess CT imaging quality.

A year later, Sen *et al.* published their own work aiming to analyse the performance of a visual-search observer for prostate SPECT images [105]. More specifically, the authors scrutinised processes for incorporating inefficiencies, i.e., background approximation, internal noise, lower thresholds, and search noise, into the VS observer. For the simulation, an extended cardiac-torso (XCAT) phantom [106] was used, and five distinct biodistributions were created for the major organs of the phantom. A total of 150 tumour locations were generated. The LROC was chosen as figure of merit, and ten distinct test strategies were studied. Furthermore, images were read by four human observers. Results showed that models were not able to reach human observer performance when applied separately, however, merging inefficiencies led to higher agreement.

Finally, Platiša *et al*. explored the impact of distinct experimental protocols on the image quality evaluation of digital pathology slides [107]. Their dataset was comprised of three images of animal pathology samples, i.e., two of gastric fundic glands and one of liver. Four nonoverlapping images were created by cropping each reference, leading to a total of twelve original stimuli. The latter were altered using nine distinct methods, i.e., adding Gaussian, unsharp masking, decreasing/increasing gamma, decreasing/adding colour saturation, adding low/high-frequency white Gaussian noise, and JPEG compression. Each alteration was only applied at a single level. Six practicing diagnostic veterinary pathologists were involved in the study, where three different protocols were compared: free-response receiver operating characteristic (FROC), DS (double stimulus), and SS (single stimulus). Under the FROC protocol, the observers were asked to mark suspected locations and to rate their confidence. Under the DS protocol, they were simultaneously presented two images and were invited to rate the similarity between the

images, as well as their preference. Finally, five criteria had to be rated under the SS protocol, i.e., perceived image quality, blur disturbance, quality of contrast, noise disturbance, and quality of colour saturation. Note that not all references and corresponding distorted versions were assessed under the three experiments. The authors chose to use the median opinion score (MdnOS) for data analysis. Thanks to the conduct of three complementary experiments, both clinical and perceived image quality, as well as similarity and preference judgments were analysed. Under the FROC protocol, pathologists rated images compressed with JPEG significantly differently than the other stimuli. Under the SS protocol, there was no statistically significant difference among alterations, except between blur and gamma. Platiša *et al*. concluded by claiming that two factors may contribute to quality scoring, i.e., the instructions given to the observers, and the context of the experiment.

5. Discussion

In this section, we scrutinise the factors influencing perceptual quality as well as the methods employed for statistical analyses in medical imaging, and we discuss the use of objective quality metrics for medical images and videos. Finally, we release recommendations for the design of future subjective experiments.

5.1. Quality's influential factors

The absence of recommendations on subjective medical image and video quality assessment shows the existence of a very complex problem: no unique optimal solution exists for all scenarios, since medical image quality is very much influenced by the contextual and the human influential factors (IFs) [108]-[109]. Context IFs include applications (e.g., diagnosis, surgery, training), clinical factors (e.g., emergency care, lesion subtleness, clinical region of interest), requirements (e.g., real-time/offline, location), medical data (clinical information (anatomical, functional physiological), acquisition modalities (e.g., ultrasound, X-ray, MRI, EEG, ECG), data types (e.g., signal, images (monochrome, colour), video (monochrome, colour)); while human IFs include expertise (e.g., years of experience, major, cultural background, educational background, pedagogical implications [47], [55]), as well as emotional state (e.g., tiredness, stress, eyestrain [110]). When context and end-users differ, the purpose of the study may require a specific test protocol. Thus, instead of giving a general recommendation, specific recommendations are needed for each combination of context IFs and human IFs, or for each category that can use the same subjective experimental protocol. The literature also lacks this type of studies.

It can be noted that several research teams compared the scores of medical professionals with the ones of naïve observers (i.e., Razaak *et al*. [42], Khan *et al*. [86], Kara *et al*. [88], Kumcu *et al*. [82], Usman *et al*. [6], and Platiša *et al*. [107]). This can be explained by the fact that the availability of medical professionals can be a limiting element for subjective experiments. In their study, Kumcu *et al*. [82] concluded that surgeons had the ability to distinguish anatomical structures, whereas naïve assessors were not sensitive to content when assessing quality. Platiša *et al*. [107] reached alike conclusions for a different medical specialty, i.e., pathology.

5.2. Data analysis methods

The International Telecommunication Union (ITU) advocates the conduct of an outlier detection and subject removal procedure on the scores obtained through a subjective image quality evaluation [16]. Such procedure is recommended as human observers may initiate doubtful scores, for instance after a misunderstanding of the instructions given by the researchers [111]. Nine of the twenty-five studies put forward this pre-processing method in an explicit manner, i.e., Chow *et al*. [2], Liu *et al*. [34], Lévêque *et al*. [3], [55], [84], Münzer *et al.* [77], Kumcu *et al.* [82], Usman *et al*. [6], and Khan *et al*. [86]. It is interesting to note that two research teams applied different methods on the raw data, i.e., an exclusion of extreme scores for Kara *et al*. [88], and a graphical technique for Platisa *et al*. [107]. No outlier exclusion procedure was mentioned by the fourteen remaining studies.

As introduced in 4.2, psychovisual experiments carried out in medical imaging present diverse requirements than for "natural" scene. Indeed, years of experience may impact participants' visual perception and, consequently, quality scoring [112]. Some studies made the choice to separate the observers according to their experience (e.g., Münzer *et al.* [77]), medical specialty (e.g., Lévêque *et al*. [3]), or even country of practice (e.g., Lévêque *et al*. [55]). Furthermore, the analysis of variance (ANOVA) can be used to analyse potential differences between participants in terms of scoring. Outtas *et al.* [47], Lévêque *et al.* [3], [55], and Liu *et al.* [34] explicitly conducted such analysis to examine the impact of participants on quality scoring.

With a view to evaluate the relationship between human scores and existing image quality metrics, most authors chose to implement correlation analyses using the Pearson linear correlation coefficient (PLCC), and Spearman's rank correlation coefficient (SROCC).

5.3. Objective quality metrics

Objective image quality assessment metrics, that can automatically predict quality perceived by human observers, are useful for real-world applications [113]. These metrics could replace subjective image quality assessment, which is expensive and cumbersome in many circumstances. In the literature, many successful objective metrics have been

developed for "natural" images and videos [114]. As one can notice from Tables I and II, a large number of authors (i.e., fourteen studies out of the twenty-five presented) decided to use quality metrics on their dataset.

In the application of perceptual objective metrics, one should be made aware of the types of these metrics. In general, objective metrics are classified into three categories, depending on the availability of the reference, i.e., distortion-free pristine image. Full-reference (FR) metrics require a full access to the reference; reduced-reference (RR) metrics make use of partial information of the reference; and no-reference (NR) metrics are not reliant on the reference [115]-[116]. As it can be seen from Tables I and II, most of the reviewed studies used FR metrics, where high-quality original images were deliberately collected and then synthetically distorted by some simulated distortions. The main goal of these studies is to measure (in the off-line scenario of image parameterisation) the impact of specific distortions on medical images. Also, FR/RR metrics may be used when "simulated" reference, e.g., phantom, is available. However, in many practical imaging scenarios, a distortion-free image is inaccessible or simply unavailable, therefore, NR metrics should be used for medical imaging quality assessment. It should be noted that developing a NR metric remains an academic challenge for natural images, and is probably even more challenging for medical images due to the wide diversity of imaging modalities. When using objective metrics for medical imaging, one should choose an appropriate type of metrics depending on the realistic clinical conditions.

There is indeed no formal or *de facto* definition of perceptual image quality for medical imaging. In the research community of natural images, there is a general consensus that image quality represents the integrated perception of the overall degree of excellence of an image. This definition is not intended to describe the utility of an image, the observers who view the image, nor the context of the acquisition process. Further research is thus needed to determine suitable definitions of perceptual image quality for medical imaging. Without the definition of "ground truth", the development of meaningful objective metrics is difficult. Different approaches have been proposed for different tasks [117]. When task-based objective metrics are applied, the studied modality and the intended task behind the quality assessment problem should first be identified. Then, an appropriate pathology to study should be chosen by considering both studied task and studied modality. According to the chosen task, an appropriate figure of merit (FOM) can be chosen. In order to design a numerical observer that can model the task performing process of medical experts, the influence of expertise should also be considered. For two decades, some works have been focusing on establishing ground truth. Personal and consensus gold standard were estimated in [118] and discussed for compressed images; effort was made in [119] to develop a gold standard using consensus-based methods; in [120], a new technique was developed for objectively evaluating quantitative nuclear imaging methods with patient data in the absence of any ground truth; and a protocol for attaining a reference diagnosis based on expert panel consensus was proposed and shown feasible in practice in [121].

Artificial intelligence (AI) algorithms have the potential to develop advanced objective quality metrics for medical imaging [122]-[123]. Building accurate AI algorithms usually requires massive training data annotated by experts, often not available in medical imaging. As a matter of fact, only three research teams out of the twenty-five studies we reviewed in this article released their quality database (i.e., Suad *et al*. [18][1], Outtas *et al*. [47][2], and Khan *et al*. [86][3]). To tackle this challenge, human-in-the-loop approaches could be explored, e.g., the machine learns and assists the experts in data annotation with reduced effort while increases understanding and reliability in the learning process. Efforts have been made to create new data with known ground truth using machine learning approaches [124].

It is worth noting that there have been significant improvements in the traditional physics-based observers. For example, a multi-template observer strategy was proposed in [125], and achieved optimal performance for detection tasks even when the signal properties were not exactly known. In addition, it should be noted that learning-based observers are still in early stages of development, while the traditional physics-based approaches are used more conventionally. With this in mind, research should focus on both strands (i.e., traditional and learning-based methods) to further improve their performance.

5.4. Recommendations

According to the results obtained by different research teams who compared medical experts with naïve observers, we recommend to carefully consider the expertise of the observers when conducting subjective quality evaluation in medical imaging. If naïve assessors are involved (e.g., in the case of pre-assessment or if no medical knowledge is required), we suggest to separately perform scores' analysis between expert, and non-expert observers.

[1] The authors did not provide the link of their database in their article, even though they mentioned a public release in their article. We tried, in vain, to contact the authors.

[2] Link to the database: http://stacks.iop.org/PMB/63/185014/mmedia

[3] Link to the database: https://drive.google.com/file/d/1SoONeacp9vvihTY7zmWssG_cnVzx16oq/view

As far as the choice of methodology is concerned, we highlighted that both single stimulus and multi stimulus have advantages and limitations. Yet, SAMVIQ [48] has the potential to integrate the benefits of both categories. Though many task-based subjective tests have been done as part of numerical observer validation experiments, their protocols are quite different. A recommendation may be necessary for a fair comparison between different laboratories. As a first step towards a recommendation, it would be useful to conduct a comparison study with these two types of test paradigms, and to explore which one is more suitable in the medical context.

In terms of statistical analyses, we strongly recommend researchers to apply a two-step pre-processing method (i.e., outlier detection and subject exclusion procedure) to the collected MOS or DMOS, as given in ITU-R recommendation BT.500-13 [16], with a view to remove any dubious scores. As for the quantitative evaluation of task-based approaches' performance, we recommend the area under the ROC curve for the detection task; while the FROC/AFROC/JAFROC curves can be used for the detection-localisation task. JAFROC1 has the highest statistical performance.

6. Conclusion

In this article, we presented a review of recent existing studies on the subjective assessment of medical image and video quality. There is a strong evidence of the significance of such a state-of-the-art, identifying available works and datasets. Furthermore, we completed our survey by discussing the methodologies used depending on several factors, like their context of application. In particular, we provided guidelines for future research works, both in terms of methodologies and data analysis. To conclude, our paper provides a better apprehension of medical imaging quality evaluation, which can be of help for researchers in the field.

References

[1] E. Krupinski, "Current perspectives in medical image perception", *Attention, Perception & Psychophysics*, vol. 72, no. 5, pp. 1205–1217, 2010.

[2] L. Chow, H. Rajagopal, and R. Paramesran, "Correlation between subjective and objective assessment of magnetic resonance images", *Magnetic Resonance Imaging*, vol. 34, no. 6, pp. 820–831, 2016.

[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] 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, no. 9, pp. 2367–2378, 2013.

[5] L. Platiša, L. Van Brantegem, Y. Vander Haeghen, C. Marchessoux, E. Vansteenkiste, and W. Philips, "Psychovisual evaluation of image quality attributes in digital pathology slides viewed on a medical color LCD display", *Medical Imaging: Digital Pathology*, vol. 8676, 2013.

[6] M. A. Usman, M. R. Usman, and S. Y. 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.

[7] The National Academies Press, *Telemedicine: A Guide to Assessing Telecommunications in Health Care*, 1996.

[8] 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.

[9] F. Schaeffel, "Processing of Information in the Human Visual System", *Handbook of Machine and Computer Vision: The Guide for Developers and Users, Second Edition*, pp. 1–33. 2007.

[10] J. Gross, F. Schmitz, I. Schnitzler, K. Kessler, K. Shapiro, B. Hommel, and A. Schnitzler, "Modulation of long-range neural synchrony reflects temporallimitations of visual attention in humans", *Proceedings of the National Academy of Sciences*, vol. 101, no. 35, pp. 13050–13055, 2004.

[11] H. Liu and Z. Wang, "Perceptual quality assessment of medical images", *Book Chapter in Encyclopedia of Biomedical Engineering*, 2018.

[12] E. A. Krupinski, and Y. Jiang, "Anniversary paper: evaluation of medical imaging systems", Medical Physics, vol. 35, no. 2, pp. 645–659, 2008.

[13] E. A. Krupinski, "Improving patient care through medical image perception research", Policy Insights from the Behavioral and Brain Sciences, vol. 2, no. 1, pp. 74–80, 2015.

[14] Recommendation ITU-R BT.1788, *Methodology for the subjective assessment of video quality in multimedia applications*, 2012.

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

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

[17] L. Lévêque, H. Liu, S. Barakovic, J. Barakovic Husic, M. Martini, M. Outtas, L. Zhang, A. Kumcu, L. Platiša, R. Rodrigues, A. Pinheiro, and A. Skodras, "On the subjective assessment of the perceived quality of medical images and videos", *10th International Conference on Quality of Multimedia Experience (QoMEX), Sardinia, Italy*, 2018.

[18] J. Suad, and W. Jbara, "Subjective quality assessment of new medical image database", *International Journal of Computer Engineering and Technology*, vol. 4, no. 5, pp. 155-164, 2013.

[19] H. Rajagopal, L. Chow, and R. Paramesran, "Subjective versus objective assessment for magnetic resonance images", *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 12, 2015.

[20] OsiriX DICOM Viewer, Available online at: https://www.osirix-viewer.com/resources/dicom-image-library/ (last accessed January 2021).

[21] H. Gudbjartsson, and S. Patz, "The Rician distribution of noisy MRI data", *Magnetic Resonance in Medicine*, vol. 34, no. 6, pp. 910-914, 1995.

[22] R. Gonzalez, and R. Woods, "Digital image processing", *Prentice Hall, 3rd Edition*, 2006.

[23] Z. Wang, A. Bovik, H. Sheik, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity", *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[24] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model", *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636-650, 2000.

[25] H. Sheikh, and A. Bovik, "Image information and visual quality", *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430-444, 2006.

[26] Alzheimer's Disease Neuroimaging Initiative, Available online at: http://www.adni.loni.usc.edu/ (last accessed November 2020).

[27] M. Sudha, and R. Sudhakar, "Two-dimensional medical image compression techniques: A survey", *International Journal on Graphics, Vision, and Image Processing*, vol. 11, no. 1, pp. 9-20, 2011.

[28] S. Sridevi, V. Vijayakuymar, and A. Ranwalage, "A survey on various compression methods for medical images", International Journal of Intelligent Systems and Applications, vol. 4, no. 3, pp. 13-19, 2012.

[29] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment", *37th IEEE Asilomar Conferences on Signals, Systems and Computers, Pacific Grove, United States of America*, vol. 2, 2003.

[30] L. Zhang, D. Zhang, and X. Mou, "FSIM: A feature similarity index for image quality assessment", *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, 2011.

[31] H. Sheikh, A. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics", *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117-2128, 2005.

[32] Z. Wang, and A. Bovik, "A universal image quality index", *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81-84, 2002.

[33] Z. Wang, and Q. Li, "Information content weighting for perceptual image quality assessment", *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185-1198, 2011.

[34] 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, no. 7, pp. 3087-3098, 2016.

[35] A. Mason, J. Rioux, S. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, and S. Beyea, "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images", *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, 2020.

[36] W. Xue, L. Zhang, X. Mou, and A. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index", *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 668-695, 2014.

[37] R. Mantiuk, K. Jim, A. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions", ACM Transactions on Graphics, vol. 30, no. 4, 2011.

[38] C. Loizou, M. Pattichis, M. Pantziaris, A. Nicolaides, and C. Pattichis, "Despeckle filtering for multiscale amplitude-modulation frequency-modulation (AM-FM) texture analysis of ultrasound images of the intima-media complex", *International Journal of Biomedical Imaging*, vol. 2014, 2014.

[39] T. Elatrozy, A. Nicolaides, T. Tegos, A. Zarka, M. Griffin, and M. Sabetai, "The effect of B-mode ultrasonic image standardisation on the echodensity of symptomatic and asymptomatic carotid bifurcation plaques," *International Angiology,* vol. 17, no. 3, pp. 179–186, 1998.

[40] C. Loizou, C. Pattichis, A. Nicolaides, and M. Pantziaris, "Manual and automated media and intima thickness measurements of the common carotid artery," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 56, no. 5, pp. 983–994, 2009.

[41] A. Mittal, R. Soudararajan, and A. Bovik, "Making a completely blind image quality analyser", *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 209-212, 2013.

[42] 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, no. 5, pp. 1552-1559, 2014.

[43] D. Chandler, and S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images", *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284-2298, 2007.

[44] M. Pinson, and S. Wolf, "A new standardized method for objectively measuring video quality", IEEE Transactions on Broadcasting, vol. 50, no. 3, pp. 312-322, 2004.

[45] M. Razaak, and M. Martini, "CUQI: Cardiac ultrasound video quality index", *Journal of Medical Imaging*, vol. 3, no. 1, 2016.

[46] M. Gray, H. Morchel, and V. Hazelwood, "Evaluating the effect of bit rate on the quality of portable ultrasound video," *IEEE 12th International Symposium on Biomedical Imaging (ISBI), New York, United States of America*, pp. 247-250, 2015.

[47] M. Outtas, L. Zhang, O. Deforges, A. Serir, W. Hamidouche, and Y. Chen, "Subjective and objective evaluations of feature selected multi output filter for speckle reduction on ultrasound images", *Physics in Medecine & Biology*, vol. 63, no. 18, 2018.

[48] F. Kozamernik, V. Steinmann, P. Sunna, and E. Wyckens, "SAMVIQ: A new EBU methodology for video quality evaluations in multimedia", *SMPTE Motion Imaging Journal*, vol. 114, no. 4, pp. 152-160, 2005.

[49] J. Kang, J. Lee, and Y. Yoo, "A new feature-enhanced speckle reduction method based on multiscale analysis for ultrasound b-mode imaging", *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1178-1791, 2016.

[50] A. Saha, and Q. Wu, "Utilizing image scales towards totally training free blind image quality assessment", *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1879-1892, 2015.

[51] M. Outtas, L. Zhang, O. Deforges, W. Hammidouche, A. Serir, and C. Cavaro-Ménard, "A study on the usability of opinion-unaware no-references natural image quality metrics in the context of medical images", *International Symposium on Signal, Image, Video and Communications (ISIVC), Tunis, Tunisia*, pp. 308-313, 2016.

[52] G. Ramos-Llorden, G. Vegas-Sanchez-Ferrero, M. Martin-Fernandez, C. Alberola-Lopez, and S. Aja-Fernandez, "Anisotropic diffusion filter with memory based on speckle statistics for ultrasound images", *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 345-358, 2014.

[53] P. Coupé, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images", *IEEE Transactions in Image Processing*, vol. 18, no. 10, pp. 2221-2229, 2009.

[54] M. Outtas, L. Zhang, O. Desforges, A. Serir, and W. Hamidouche, "Multi-output speckle reduction filter for ultrasound medical images based on multiplicative multiresolution decomposition", *IEEE International Conference on Image Processing (ICIP), Beijing, China*, 2017.

[55] L. Lévêque, W. Zhang, and H. Liu, "International comparison of radiologists' assessment of the perceptual quality of medical ultrasound video", *11th International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany*, 2019.

[56] The Cancer Imaging Archive, Available online at https://www.cancerimagingarchive.net/ (last accessed January 2021).

[57] L. Tang, C. Tian, L. Li, B. Hu, W. Yu, and K. Xu, "Perceptual quality assessment for multimodal medical image fusion", *Signal Processing: Image Communication*, vol. 85, 2020.

[58] S. Liu, T. Zhang, H. Li, J. Zhao, and H. Li, "Medical image fusion based on nuclear norm minimization", *International Journal of Imaging Systems and Technology*, vol. 25, no. 4, pp. 310-316, 2015.

[59] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation", *Information Fusion*, vol. 24, pp. 147-164, 2015.

[60] R. Shen, I. Cheng, and A. Basu, "Cross-scale coefficient selection for volumetric medical image fusion", *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1069-1079, 2013.

[61] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering", *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864-2875, 2013.

[62] S. Das, and M. Kundu, "NSCT-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency", *Medical and Biological Engineering and Computing (MBEC)*, vol. 50, no. 10, pp. 1105-1114, 2012.

[63] S. Liu, J. Zhao, and M. Shi, "Medical image fusion based on improved sum modified Laplacian", *International Journal of Imaging Systems and Technology*, vol. 25, no. 3, pp. 206-212, 2015.

[64] Y. Liu, X. Chen, and R. Ward, "Image fusion with convolutional sparse representation", *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1882-1886, 2016.

[65] L. Tang, J. Qian, L. Li, J. Hu, and X. Wu, "Multimodal medical image fusion based on discrete Tchebichef moments and pulse coupled neural network", *International Journal of Imaging Systems and Technology*, vol. 27, no. 1, pp. 57-65, 2017.

[66] C. Xydeas, and V. Petrovic, "Objective image fusion performance measure", *Electronics Letters*, vol. 36, no. 4, pp. 308-309, 2000.

[67] C. Yang, J. Zhang, and X. Wang, "A novel similarity-based quality metric for image fusion", *Information Fusion*, vol. 9, no. 2, pp. 156-160, 2008.

[68] S. Li, J. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency", *Information Fusion*, vol. 2, no. 3, pp. 169-176, 2001.

[69] J. Zhao, R. Laganiere, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement", *International Journal of Innovative Computing, Information and Control*, vol. 3, no. 6, pp. 1433-1447, 2007.

[70] Y. Zheng, E. Essock, B. Hansen, and A. Haun, "A new metric based on extended spatial frequency and its application to DWT based fusion algorithms", *Information Fusion*, vol. 8, no 2., pp. 177-192, 2007.

[71] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion", *Electronics Letters*, vol. 38, no. 7, pp. 313-315, 2002.

[72] K. Ma, Z. Duanmu, H. Yeganeh, and Z. Wang, "Multi-exposure image fusion by optimizing a structural similarity index", *IEEE Transactions on Computing Imaging*, vol. 4, no. 1, pp. 60-72, 2018.

[73] A. Sholehkerdar, J. Tavakoli, and Z. Liu, "In-depth analysis of Tsallis entropy-based measures for image fusion quality assessment", *Optical Engineering*, vol. 58, no. 3, pp. 1-16, 2019.

[74] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation", *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508-517, 2018.

[75] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks", IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1202-1213, 2018.

[76] N. Nouri, D. Abraham, J-M. Moureaux, M. Dufaut, J. Hubert, and M. Perez, "Subjective MPEG2 compressed video quality assessment: Application to tele-surgery", *IEEE 7th International Symposium on Biomedical Imaging (ISBI), Rotterdam, Netherlands*, 2010.

[77] B. Münzer, K. Schoeffmann, L. Böszörmenyi, J. Smulders, and J. Jakimowicz, "Investigation of the impact of compression on the perceptional quality of laparoscopic videos", *IEEE 27th International Symposium on Computer-Based Medical Systems, New York, USA*, 2014.

[78] A. Chaabouni, Y. Gaudeau, J. Lambert, J. Moureaux, and P. Gallet, "Subjective and objective quality assessment for H264 compressed medical video sequences", *IEEE 4th International Conference on Image Processing Theory, Tools and Applications (IPTA), Paris, France*, 2014.

[79] A. Chaabouni, Y. Gaudeau, J. Lambert, J. Moureaux, and P. Gallet, "H.264 medical video compression for telemedicine: A performance analysis", *Innovation and Research in Biomedical Engineering (IRBM)*, vol. 37, no. 1, pp. 40-48, 2015.

[80] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "Two new full-reference quality metrics based on HVS", *Workshop on Video Processing and Quality Metrics (VPQM), Scottsdale, United States of America*, 2006.

[81] A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain", *IEEE Transactions on Image Processing*, vol. 20, no. 2, pp. 209-212, 2013.

[82] A. Kumcu, K. Bombeke, H. Chen, L. Jovanov, L. Platiša, H. Luong, J. Van Looy, Y. Van Niewenhove, P. Schelkens, and W. Philips, "Visual quality assessment of H.264/AVC compressed laparoscopic video", *SPIE Image Perception, Observer Performance, and Technology Assessment*, vol. 9037, 2014.

[83] A. Kumcu, K. Bombeke, L. Platiša, 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, no. 1, pp. 48-63, 2017.

[84] L. Lévêque, H. Liu, C. Cavaro-Ménard, Y. Cheng, and P. Le Callet, "Video quality perception in telesurgery", *IEEE 19th International Workshop on Multimedia and Signal Processing (MMSP), Luton, United Kingdom*, 2017.

[85] MSU Perceptual Video Quality Tool, Available online at http://compression.ru/video/quality_measure/ (last accessed January 2021).

[86] Z. Khan, A. Beghdadi, F. Cheikh, M. Kaaniche, E. Pelanis, R. Palomar, A. Fretland, B. Edwin, and O. Elle, "Towards a video quality assessment-based framework for enhancement of laparoscopic videos", *SPIE Image Perception, Observer Performance, and Technology Assessment*, vol. 11316, 2020.

[87] A. Mittal, M. Saad, and A. Bovik, "A completely blind video integrity oracle", *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 289-300, 2015.

[88] P. Kara, P. Kovacs, S. Vagharshakyan, M. Martini, S. Imre, A. Barsi, K. Lackner, and T. Balogh, "Perceptual quality of reconstructed medical images on projection-based light field displays", *EAI International Conference on Mobile Medical Multimedia Technologies, Applications and Services (M3Apps), Budapest, Romania*, 2016.

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

[90] D. Rouse, R. Pepion, P. Le Callet, and S. Hemami, "Tradeoffs in subjective testing methods for image and video quality", *SPIE Human Vision and Electronic Imaging, San Jose, United States of America*, 2010.

[91] L. Zhang, C. Cavaro-Ménard, and P. Le Callet, "Key issues and specificities for the objective medical image quality assessment", *6th International Workshop on Video Processing and Quality Metrics (VPQM), Scottsdale, United States,* 2012.

[92] N. Obuchowski, "ROC analysis", *Fundamentals of Clinical Research for Radiologists*, vol. 184, no. 2, pp. 364-372, 2005.

[93] C. Metz, "ROC analysis in medical imaging: A tutorial review of the literature", *Radiological Physics and Technology*, vol. 1, no. 1, pp. 2-12, 2008.

[94] M. Kalayeh, T. Marin, and J. Brankov, "Generalization evaluation of machine learning numerical observers for the image quality assessment", *IEEE Transactions on Nuclear Science*, vol. 60, no. 3, pp. 1609-1618, 2013.

[95] I. Lorente, and J. Brankov, "Active learning for image quality assessment by model observer", *11th IEEE International Symposium on Biomedical Imaging (ISBI), Beijing, China*, 2014.

[96] T. Marin, M. Kalayeh, F. Parages, and J. Brankov, "Numerical surrogates for human observers in myocardial motion evaluation from SPECT images", *IEEE Transactions on Medical Imaging*, vol. 33, no. 1, pp. 38-47, 2013.

[97] P. Pretorius, M. King, B. Tsui, K. Lacroix, and W. Xia, "A mathematical model of motion of the heart for use in generating source and attenuation maps for simulating emission imaging", *Medical Physics*, vol. 26, no. 11, pp. 2323-2332, 1999.

[98] G. Wen, M. Markey, T. Haygood, and S. Park, "Model observer for assessing digital breast tomosynthesis for multi-lesion detection in the presence of anatomical noise", *Physics in Medicine and Biology*, vol. 63, no. 4, 2018.

[99] D. Chakraborty, "A brief history of free-response receiver operating characteristic paradigm data analysis", *Academic Radiology*, vol. 20, no. 7, pp. 915-919, 2013.

[100] W. Zhou, H. Li, and M. Anastasio, "Approximating the ideal observer and Hotelling observer for binary signal detection tasks by use of supervised learning methods", *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2456-2468, 2019.

[101] S. He, W. Zhou, H. Li, and M. Anastasio, "Learning numerical observers using unsupervised domain adaptation", *SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, *Houston, United States of America*, 2020.

[102] L. Zhang, C. Cavaro-Ménard, P. Le Callet, and J-Y. Tanguy, "A perceptually relevant channelized joint observer for the detection-localization of parametric signals", *IEEE Transactions on Medical Imaging*, vol. 31, no. 10, pp. 1875-1888, 2012.

[103] T. Xu, L. Zhang, Y. Chen, H. Shu, and L. Luo, "Quality assessment based on PCJO for low-dose CT images", *14th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine (Fully3D), Xi'an, China*, 2017.

[104] S. Leng, L. Yu, Y. Zhang, R. Carter, A. Toledano, and C. McCollough, "Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain", *Medical Physics*, vol. 40, no. 8, 2013.

[105] A. Sen, F. Kalantari, and H. Gifford, "Impact of anatomical noise on model observers for prostate SPECT", *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Seattle, United States of America*, pp. 1-6, 2014.

[106] W. Segars, and B. Tsui, "MCAT to XCAT: The evolution of 4D computerized phantoms for imaging research", *Proceedings of the IEEE Institute of Electrical and Electronical Engineers*, vol. 97, no. 12, pp. 1954-1968, 2013.

[107] L. Platiša, L. Van Brantegem, A. Kumcu, R. Ducatelle, and W. Philips, "Influence of study design on digital pathology image quality evaluation: The need to define a clinical task", *Journal of Medical Imaging*, vol. 4, no. 2, 2017.

[108] 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, Prague, Czech Republic*, 2012.

[109] 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.

[110] C. Cavaro-Ménard, L. Zhang, and P. Le Callet, "QoE for telemedicine: Challenges and trends", *SPIE Optics and Photonics, Applications of Digital Image Processing, San Diego, United States of America*, vol. 8856, 2013.

[111] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment", *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478-2491, 2012.

[112] E. Krupinski, and R. Weinstein, "Changes in visual search patterns of pathology residents as they gain experience", *SPIE Image Perception, Observer Performance, and Technology Assessment, Lake Buena Vista, United States*, vol. 7966, 2011.

[113] Z. Wang, "Applications of objective image quality assessment methods", *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 137-142, 2011.

[114] Z. Wang, and A. Bovik, "Modern image quality assessment", *Syntheses Lectures on Image, Video and Multimedia Processing,* Morgan & Claypool Publishers, 2006.

[115] Z. Wang, and A. Bovik, "Reduced and no-reference image quality assessment", *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29-40, 2011.

[116] L. Atidel, C. Larabi, A. Beghdadi, E. Viennet, and A. Bouridane, "Knowledge-based taxonomic scheme for full-reference objective image quality measurement models", *Journal of Imaging Science and Technology*, vol. 60, no. 6, pp. 1-15, 2016.

[117] H. Barrett, and K. Myers, "Image quality", *Foundations of Image Science, Chapter 14, Wiley,* 2003.

[118] P. Cosman, R. Gray, and R. Olshen "Quality evaluation for compressed medical images: Diagnostic accuracy", *Handbook of Medical Imaging: Processing and Analysis Management, Chapter 6*, pp. 821-839, 2000.

[119] D. Miller, K. O'Shaughnessy, S. Wood, and R. Castellino, "Gold standards and expert panels: A pulmonary nodule case study with challenges and solutions", *Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, vol. 5372, 2004.

[120] A. Jha, B. Caffo, and E. Frey, "A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods", *Physics in Medicine & Biology*, vol. 61, no. 7, pp. 2780-2800, 2016.

[121] R. Handels, C. Wolfs, P. Aalten, P. Bossuyt, M. Joore, A. Leentjens, J. Severens, and F. Verhey, "Optimizing the use of expert panel reference diagnoses in diagnostic studies of multidimensional syndromes", BMC Neuroloy, vol. 14, no. 1, 2014.

[122] Z. Yu, M. Rahman, T. Schindler, R. Gropler, R. Laforest, R. Wahl, and A. Jha, "AI-based methods for nuclear-medicine imaging: Need for objective task-specific evaluation", *Journal of Nuclear Medicine*, vol. 61, pp. 575, 2020.

[123] K. Li, W. Zhou, H. Li, and M. Anastasio, "Task-based performance evaluation of deep neural network-based image denoising", *SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, vol. 11599, pp. 114-118, 2021.

[124] M. Willemink, W. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. Folio, R. Summers, D. Rubin, and M. Lungren, "Preparing medical imaging data for machine learning", *Radiology*, vol. 295, no. 1, 2020.

[125] X. Li, A. Jha, M. Ghaly, F. Elsahaby, J. Links, and E. Frey, "Use of sub-ensembles and multi-template observers to evaluate detection task performance for data that are not multivariate normal", *IEEE Transactions on Medical Imaging*, vol. 36, no. 4, pp. 917-929, 2018.