THÈSE

pour obtenir le grade de

Docteur de l'université Sorbonne Paris Nord

Discipline : "Doctorat de Ingenierie Informatique"

présentée et soutenue publiquement par

Zohaib Amjad KHAN

le 29 Janvier 2021

Learning based quality assessment for medical imaging in the context of liver cancer treatment

Directeur de thèse : **Prof. Azeddine Beghdadi**

Co-directeur de thèse : **Prof. Faouzi Alaya Cheikh**

Co-directeur de thèse : **Dr. Mounir Kaaniche**

JURY

Karen Eguiazarian,	Professeur, Tampere University, Finland	Rapporteur
Lu Zhang,	Maître de conférences (HDR), INSA, Rennes	Rapporteur
Catalin Fetita,	Professeur, ARTEMIS, Telecom SudParis	Rapporteur
Frédéric Dufaux,	Directeur de recherche CNRS, Université Paris-Saclay	Examinateur
Azeddine Beghdadi,	Professeur, USPN	Examinateur
Faouzi Alaya Cheikh,	Professeur, NTNU, Norway	Examinateur
Mounir Kaaniche,	Maître de conférences (HDR), USPN	Examinateur

To

My Dearest Mother,

for her infinite kindness, love and guidance.

Résumé

Le cancer du foie est le quatrième cancer le plus mortel au monde avec un taux de récidive élevé. Son traitement implique souvent une ablation ou une résection chirurgicale du tissu affecté. Pour améliorer les résultats de cette procédure, de nouvelles méthodes intègrent des interventions chirurgicales guidées par l'image, qui exploite différentes modalités d'imagerie médicale. Parmi ces modalités, on peut citer l'échographie, la tomographie (CT), l'imagerie par résonance magnétique (IRM) et la vidéo laparoscopique mono-vue (2D) ou stéréoscopique (3D).

L'évaluation de la qualité de ces images et vidéos médicales est extrêmement critique aussi bien pour la fiabilité du diagnostic médical que pour la précision de la chirurgie. Cependant, l'exploitation des diverses modalités d'imagerie médicale pour l'évaluation de la qualité d'image est une tâche assez complexe. De plus, les performances de la chirurgie guidée par l'image, sont tributaires de la qualité de différentes tâches critiques telles que le recalage d'images, l'amélioration de la qualité et la segmentation d'image. Cette thèse est consacrée principalementà la recherche de solutions efficaces pour répondre à ces deux problématiques critiques, à savoir l'évaluation de la qualité de l'image dans le contexte médical et l'évaluation des performances des algorithmes de traitement d'images utilisés dans le diagnostic et la chirurgie guidée par l'image.

Dans cette thèse, différentes méthodes d'évaluation de la qualité d'image issues des trois modalités d'imagerie médicale, à savoir les vidéos laparoscopiques 2D, les images laparoscopiques 3D et la tomographie sont proposées et évaluées. Plus précisément, pour les images et vidéos laparoscopiques mono-vue et stéréoscopiques, nous nous sommes focalisés sur l'évaluation de la qualité des images/vidéos brutes affectées par des distorsions. Pour les images CT, une nouvelle approche pour l'évaluation de la qualité d'image ainsi et que l'estimation du niveau de rehaussement de qualité est proposée.

Une méthode préliminaire portant sur l'évaluation objective de la qualité d'images stéréoscopiques et basée sur des modèles statistiques exploitant les corrélations inter-vues a été développée et évaluée sur une base publique d'images stéréoscopiques de scènes naturelles en raison de l'absence de bases d'images médicales annotées.

Pour parer à ce manque de données médicales, nous avons construit une nouvelle base de

données à partir de séquences de vidéos laparoscopiques (2D) contenant quelques distorsions types à différents niveaux de sévérité. Cette base contient les résultats de tests psycho-visuels effectués par un panel d'experts du milieu médical et ainsi que de observateurs non experts. Nous avons aussi développé une méthode de classification des distorsions types dans les images laparoscopiques 2D basée sur l'apprentissage profond. Cette méthode a été étendue pour l'évaluation de la qualité objective sans référence des vidéos laparoscopiques 2D qui s'est révélée plus efficace que les méthodes comparables de l'état de l'art.

Enfin, pour l'évaluation de l'amélioration du contraste des images CT, nous avons d'abord identifié les critères le plus pertinents permettant de quantifier le niveau d'amélioration. Une nouvelle stratégie basée sur l'apprentissage automatique en combinant différentes mesures d'évaluation de qualité de l'amélioration du contraste a été ainsi proposée. Dans cette méthode, les étapes d'apprentissage et de validation reposent sur les mesures de performance de la segmentation des images. Les résultats obtenus montrent une amélioration substantielle par rapport aux méthodes existantes.

Les multiples contributions de cette thèse à différents niveaux ont permis de faire avancer un sujet assez complexe en raison du manque de données médicales annotées et de l'absence de consensus quant aux métriques de qualité d'images dans le contexte de l'imagerie médicale. Ce travail a aussi mis l'accent sur différentes pistes qui méritent d'être approfondies. Il est important de noter que l'évaluation de la qualité dans ce contexte ne se limite pas seulement à l'évaluation de la qualité des images médicales de différentes modalités, mais concerne également l'évaluation des méthodes de traitement d'images médicales telles que l'amélioration, la segmentation et le recalage d'images. De plus, avec l'émergence et l'efficacité des techniques d'apprentissage profond, le besoin de grandes bases de données pour l'évaluation de la qualité avec l'implication des experts médicaux est maintenant devenu une exigence essentielle qui doit être satisfaite.

Mots Clés— imagerie médicale, évaluation de la qualité d'image, évaluation objective et subjective, apprentissage en profondeur, évaluation de l'amélioration du contraste.

Abstract

Liver cancer is the fourth deadliest cancer in the world with a high rate of recurrence. Its treatment often involves ablation or surgical resection of the affected tissue. To improve the outcomes of this procedure, new methods incorporate image-guided surgical interventions, which require input from multiple imaging modalities. In the case of liver cancer treatment, these modalities are the Ultrasound, the Computed Tomography (CT) imaging, the Magnetic Resonance Imaging (MRI), monoscopic (2D) and stereoscopic (3D) laparoscopic videos.

Quality assessment of these medical images and videos is extremely critical, throughout the diagnosis and treatment phases, not only for an improved diagnosis but also for an error-free performance of the surgery. However, multiple modalities with different characteristics like CT, MRI, ultrasound and laparoscopic videos make the quality assessment a challenging task. Moreover, in image-guided surgery, there are different image processing tasks involved like registration, enhancement and segmentation whose performance need to be evaluated. This thesis is mainly focused on finding solutions to both of these two critical problems, namely medical image quality assessment and performance evaluation of imaging tasks used in image-guided interventions.

In this thesis, we have proposed image quality assessment methods for three of the most important imaging modalities of image-guided surgery namely 2D laparoscopic videos, 3D laparoscopic images and the CT. More specifically, for monoscopic and stereoscopic laparoscopic images/videos, we have focused on quality assessment of the non-processed images/videos affected by distortions, whereas for CT images we have proposed a new method for quality evaluation of their enhancement.

First of all, we have developed a no-reference objective image quality assessment method for stereoscopic images based on joint statistical features. However, due to the lack of labeled 3D laparoscopic data, we could not validate the results of this method for medical images, and so, a standard natural stereo image dataset has been used for evaluation purpose.

We have then taken on the issue of the lack of labeled data and have constructed a new 2D laparoscopic video quality database. Evaluations from both medical expert and non-expert

observers have been included in this database. Thereafter we have developed an effective deep learning method for simultaneous distortion classification and ranking for 2D laparoscopic images. We have then further extended this method for no-reference objective quality assessment of 2D laparoscopic videos and have obtained the best results as compared to other state-of-the-art methods.

For the evaluation of contrast enhancement in CT, we have first identified the most important criteria of enhancement and the appropriate metrics to quantify these criteria. Thereafter, we have proposed a novel goal-oriented machine learning-based strategy to combine these different contrast enhancement evaluation (CEE) metrics. For training and validation of our method, we have used the performance evaluation scores of the subsequent task, namely segmentation, as the labels. The results of our proposed method show a substantial improvement compared to state-of-the-art CEE metrics.

Despite the multiple contributions in this thesis for medical image quality assessment and enhancement evaluation, there is still a considerable gap to be filled in this field. It is important to note that assessment of quality in this context is not only restricted to the evaluation of the quality of medical images from different modalities, but also encompasses the evaluation of processing tasks like enhancement, segmentation and registration. Furthermore, with the rise of deep learning, the need of large databases for quality assessment with input from medical experts, has now become an essential requirement that yet needs to be filled.

Keywords— medical imaging, image quality assessment, video quality assessment, objective and subjective assessment, statistical modeling, deep learning, contrast enhancement evaluation, segmentation, registration, guided, intervention, laparascopic.

Contents

List of Tables

List of Figures

Abbreviations

AMBE	Absolute Mean Brightness Error
AME	Absolute Measure of Enhancement
AMEE	Absolute Measure of Enhancement by Entropy
BGG	Bivariate Generalized Gaussian
BerGG	Bernoulli Generalized Gaussian
BRISQUE	Blind/referenceless Image Spatial Quality Evaluator
CEE	Contrast Enhancement Evaluation
CEED	Contrast Enhancement Evaluation Database
CHO	Channelized Hotelling Observer
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CT	Computed Tomography (CT)
DCT	Discrete Cosine Transform
DE	Discrete Entropy
DNN	Deep Neural Network
EC	Edge Content
EME	Measure of Enhancement
EMEE	Measure of Enhancement by Entropy
FR-IQA	Full-Reference Image Quality Assessment
GHE	Global Histogram Equalization
HO	Hotelling Observer
HVS	Human Visual System
IEM	Image Enhancement Measure
IQA	Image Quality Assessment
KROCC	Kendall Rank-Order Correlation Coefficient

LOE	Lightness Order Error
LOM	Lightness Order Measure
LVQ	Laparoscopic Video Quality
MAE	Mean Absolute Error
MCCEE	Multi-criteria Contrast Enhancement Evaluation
MGG	Multivariate Generalized Gaussian
MRETINEX	Multi-scale Retinex
MRI	Magnetic Resonance Imaging
NR-IQA	No-Reference Image Quality Assessment
NSS	Natural Scene Statistics
OR	Outlier Ratio
PLCC	Pearson Linear Correlation Coefficient
PSNR	Peak Signal-to-Noise Ratio
ResNet	Residual Network
RMSC	Root Mean Square Contrast
RMSE	Root Mean Square Error
RSE	Radial Spectral Energy
RR-IQA	Reduced-Reference Image Quality Assessment
SDME	Second Derivative-like Measurement
SIQA	Stereoscopic Image Quality Assessment
SMO	Structure Measure Operator
SROCC	Spearman Rank-Order Correlation Coefficient
SSIM	Structural Similiarity Index Measure
SVM	Support Vector Machine
SVR	Support Vector Regression
VIF	Visual Information Fidelity
VQA	Video Quality Assessment

Notations

(m,n)	pixel coordinates
$I(m,n)$	image value at pixel coordinates (m,n)
$F_i(m,n)$	original video frame value at time i and pixel coordinates (m,n)
$d_i(m,n)$	distorted video frame value at time i and pixel coordinates (m,n)
$I^{(v)}$	left $v=l$ or right $v=r$ view from stereo pair
$\mathbb{R}$	set of real numbers
Γ	Gamma function
Σ_j	scatter matrix at resolution j
$E(.)$	expected value or mean
$sign(.)$	signum function
$\mathbf{f}$	feature vector
${}^{L}\mathbf{C}_2$	combination of 2 numbers out of L
$A*B$	convolution of A and B
$\|\|.\|\|$	norm operator
$\mathcal{L}_X$	Loss function of type $X=\{\mathcal{MSE},\mathcal{CE}\}$

CHAPTER 1

Introduction

1.1 Context and Motivation

With the advancement in medical image acquisition technology in the recent years, different new modalities have appeared. This has not only resulted in enhancing diagnostic capabilities but has also inspired their use during surgery like in laparoscopic, robotic and image-guided surgeries (Figure 1.1). Moreover, the use of video recording of preoperative and intraoperative procedures has also recently become more prevalent because of its advantage in allowing accurate decision-making in Minimally Invasive Surgery (MIS) [1]. Examples of commonly used medical modalities are X-rays, Magnetic Resonance Imaging (MRI), Computerized Tomography (CT), Positron Emission Tomography (PET), Single-Photon Emission Computed Tomography (SPECT), Ultrasound, infrared, fluorescent, microwave and microscopic imaging. Some application specific modalities also exist like Mammography for breast cancer and Trans-rectal ultrasound (TRUS) for prostate brachytherapy dosimetry [2]. Besides these, some other modalities like stereoscopic imaging and DynaCT provide 3D views of the target area.

Use of one or more of these modalities in cancer treatment before, during and after surgical interventions is an established practice now. For instance, with the help of image-guided systems, surgeons can make use of pre-operative planning to navigate their instruments to the accurate location of tumor. However, there still exists a considerable gap in a streamlined and efficient use of surgical navigation technologies for certain kinds of cancers which, if tackled, could improve the survival prognosis for the affected patients. Liver cancer is one of them. Being the fourth

(a) Laparoscopic surgery **(b)** Image-guided surgery **(c)** Robotic surgery

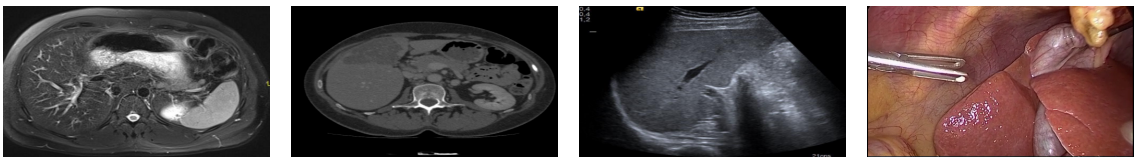
Figure 1.1: Modern surgeries that use imaging modalities

most common cause of cancer mortality, improvements in surgical navigation and management of liver cancers is an important need of today.

Liver cancer may occur either in the form of hepatocellular carcinoma (HCC) or as a result of metastases from other cancers like colorectal. The treatment of liver cancer often involves ablation or surgical resection of the affected tissue. However, eligibility of patients for resection of tumour is currently low and cancer recurrence rate is high. To improve both these parameters, surgical workflow for liver cancer resection which includes pre-operative planning, intra-operative resection navigation and ablation monitoring and post-operative quality control, needs to be improved.

In the case of liver cancer treatment, the medical modalities commonly used are the Ultrasound, the Computed Tomography (CT) imaging, the Magnetic Resonance Imaging (MRI), monoscopic (2D) laparoscopy and stereoscopic (3D) laparoscopy (Figure 1.2). The image from each modality may provide some unique and specific information. The images from multiple modalities are sometimes combined to create a composite image containing information from the input modalities in a process called image fusion. Throughout the treatment phase, it is extremely critical to have a methodology to keep a check on the image quality during each step, starting from acquisition, to ensure that the useful information is not lost.

Image quality is commonly defined in terms of visible distortions in an image like color shifts, blurriness, noise and blockiness [3]. Image quality can be assessed either subjectively



(a) MRI **(b)** CT image **(c)** Ultrasound **(d)** Laparoscopic image

Figure 1.2: Medical imaging modalities used for liver cancer surgery

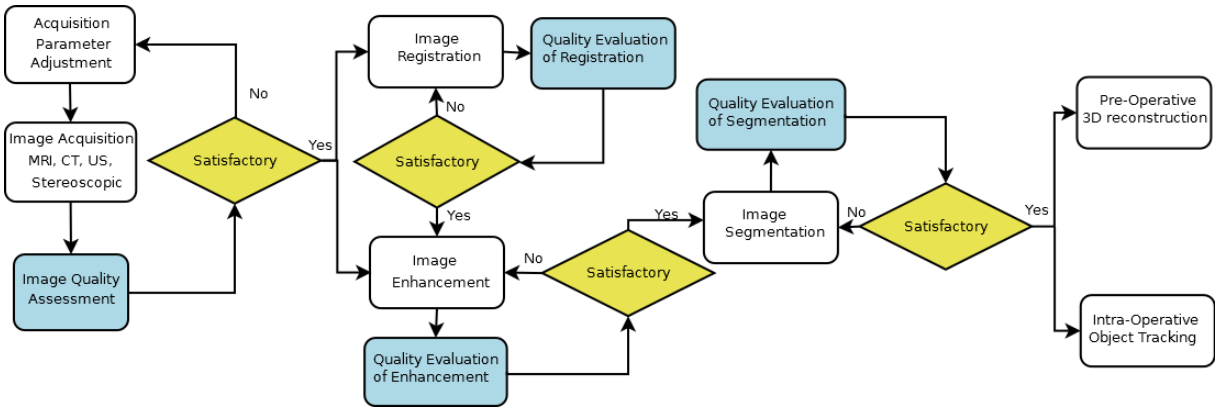


Figure 1.3: General Image Processing Pipeline for Surgical Navigation

using human feedback or objectively. However, the use of human feedback is time-consuming and prone to inconsistency and may not always be feasible in real world applications such as video-guided surgery. Hence, some efficient algorithms are required for objective image quality assessment. Moreover, it is also useful to have evaluation metrics for assessing the results and performance of other steps in the processing pipeline like image registration, enhancement and segmentation. This research work aims to investigate and come up with effective solutions to both these important issues. To illustrate how quality assessment fits in the overall setup, Figure 1.3 shows a flowchart depicting a generic image processing pipeline for surgical navigation. The highlighted steps indicate quality assessment of images and different processing steps.

1.2 Thesis Objectives

It is evident from the existing literature that there is still a wide gap that needs to be filled in image quality assessment for multimodal medical imaging. Moreover, not much work has been done to model accurately the diagnostic quality of medical images as perceived by the experts in the medical field. In the prospect of planning and navigation for liver resection surgery, both these aspects are fundamental to the overall success of the procedure. With the limited resources available for research on medical image quality assessment, this thesis focuses on finding solutions to the following problems.

1.2.1 Image and Video Quality Assessment of Laparoscopic imaging

This part of the thesis focuses on developing a medical database for laparoscopic videos consisting of corresponding subjective scores as well. The challenge is to use an effective way of scoring, involving both medical experts, for diagnostic quality assessment, and non-experts for perceptual

quality assessment. Moreover in this part, another main focus is the development of efficient objective quality assessment methods for image, video and stereoscopic images. The important questions that have been investigated are:

- Creating a standard database of laparoscopic videos for liver surgery which contains all relevant distortions and artifacts that can affect the pre-operative diagnosis or intra-operative performance. The database would also contain subjective quality scores.
- Finding out a methodology to effectively classify distortions in the images and videos.
- Developing efficient deep learning based and machine learning based metrics for quality assessment of laparoscopic images and videos

1.2.2 Contrast Enhancement Evaluation of CT images

The second part of the research mainly focuses on the development of efficient evaluation methods to judge the quality of contrast enhancement in CT images. The questions relevant to this part are:

- Designing metrics which are able to assess the quality of enhancement of medical images from specific modalities.
- Incorporating methodology to detect over-enhancement artifacts besides detecting improvement in contrast.

1.3 Thesis Contributions

The main contributions of this research can be summarized as follows:

- Development of different objective quality assessment methods for image modalities commonly used in image-guided surgery.
- For assessment of stereoscopic images, exploitation of inter-view dependencies and intra-view spatial redundancies of the stereo wavelet representations by resorting to a joint statistical modeling.
- Construction of a new quality database (LVQ) with subjective scores for 2D laparoscopic videos that has been made publicly available.

- Synthesis of distorted laparoscopic videos for LVQ database with some common distortions as well as with some uncommon ones like uneven illumination using a specialized mask and smoke using screen blending.
- Inclusion of subjective evaluations from expert surgeons for the new LVQ database.
- Exploitation of deep learning for classifying distortions in a laparoscopic image and estimating their severity level.
- Proposal of a new video quality assessment metric for 2D laparoscopic videos based on a combination of residual network and fully-connected network.
- Identification of the important criteria and metrics for an effective contrast enhancement method.
- Proposition of a new learning based metric for contrast enhancement evaluation of natural and CT images using machine learning paradigm.
- Use of a task-based approach for evaluating contrast of CT images based on performance of the subsequent segmentation task.
- Proposal of a new metric called Luminance Mean to Range (LMR) measure for detecting and evaluating the severity of uneven illumination in an image.

1.4 Thesis Outline

The thesis is divided into multiple chapters. Chapter 2 first gives a literature review on image and video quality assessment. In this chapter, we have described existing state-of-the-art methods in subjective and objective quality assessment of different imaging modalities, as well as various benchmarking methods for evaluation of assessment methods. Each of the remaining chapters discusses in detail a separate contribution to the thesis topic. Figure 1.4 illustrates the different contributions arranged according to the different modalities involved and sorted by the order in which they appear in this thesis.

In Chapter 3, we first present a novel no-reference image quality assessment method for stereoscopic image modality. More specifically, we highlight the use of joint statistics for sterescopic image quality assessment in this chapter and illustrate some performance improvements compared to the state-of-the-art.

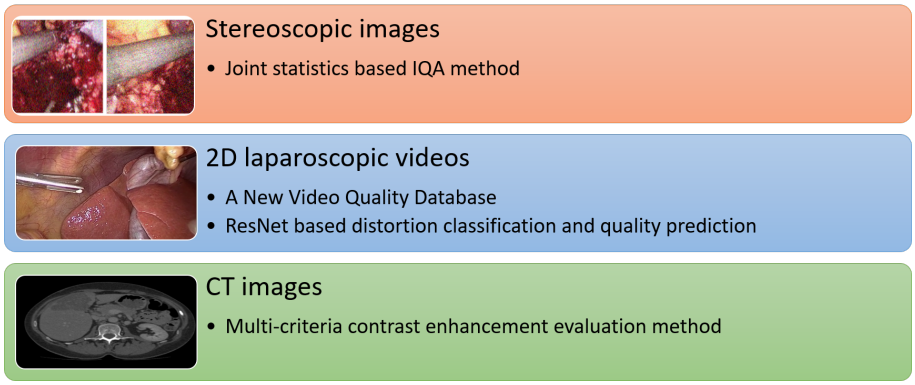


Figure 1.4: Thesis contributions arranged by different modalities and ordered by chapters.

In Chapter 4 we present our next contribution. It relates to the construction of a new subjective video quality database for monoview laparoscopic videos. Here, we have discussed in detail the procedure for development of distorted videos, subjective evaluations and objective assessment.

Chapter 5 presents our next work in the scope of this thesis where we have proposed a novel no-reference deep learning based method for quality assessment of laparoscopic videos. In this chapter, we further analyse the performance of our method in comparison to other video quality methods.

Our last contribution is detailed in Chapter 6 where we have presented a new and effective solution to contrast enhancement evaluation of images with special emphasis on CT modality. Here, unlike existing methods, we illustrate the use of carefully selected multiple criteria for evaluation of contrast enhancement in images.

Finally, in Chapter 7 we provide some conclusions from the achieved work and some perspectives for the future work.

1.5 List of Publications

Based on the research work presented in this thesis, some papers have been published and submitted for publication in international journals and conferences as following:

Journal papers:

1) **Khan, Z. A.**, Beghdadi, A., Kaaniche, M., Alaya Cheikh, F. and Gharbi O., "End-to-end Blind Quality Assessment for Laparoscopic Videos using Neural Networks", IEEE Transactions on Medical Imaging (Submitted).

2) Naseem, R., **Khan, Z. A.**, Satpute, N., Beghdadi, A., Alaya Cheikh, F. and Olivares, J., "Cross-modality guided liver CT contrast enhancement for improved tumor segmentation",

IEEE Access (Submitted).

Conference papers:

1) **Khan, Z. A.**, Beghdadi, A., Kaaniche, M. and Alaya Cheikh, F., 2020, October. "Residual Networks Based Distortion Classification and Ranking for Laparoscopic Image Quality Assessment". In 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, October 2020, pp. 176-180. IEEE.

2) **Khan, Z. A.**, Beghdadi A., Alaya Cheikh, F., Kaaniche, M. and Qureshi, M. A. "A Multi-Criteria Contrast Enhancement Evaluation Measure using Wavelet Decomposition". In 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, September 2020, pp. 1-6. IEEE.

3) **Khan, Z. A.**, Beghdadi, A., Alaya Cheikh, F., Kaaniche, M., Pelanis, E., Palomar, R., Fretland, Å. A., Edwin, B. and Elle, O. J. "Towards a video quality assessment based framework for enhancement of laparoscopic videos". In Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, Houston, Texas USA, February 2020, Vol. 11316, p. 113160P. International Society for Optics and Photonics, SPIE.

4) **Khan, Z. A.**, Kaaniche, M., Beghdadi, A. and Alaya Cheikh, F., 2018, November. "Joint statistical models for no-reference stereoscopic image quality assessment". In 2018 7th European Workshop on Visual Information Processing (EUVIP), Tampere Finland, November 2018, pp. 1-5. IEEE.

Challenge session organization:

- "Real-time distortion classification in laparoscopic videos". Organized by Beghdadi, A., **Khan, Z. A.**, Alaya Cheikh, F., Kaaniche, M., Pelanis, E., Palomar, R., Fretland, Å. A., Edwin, B. and Elle, O. J. In 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, October 2020.

CHAPTER 2

State-of-the-Art

2.1 Overview

In this chapter, we discuss in detail the state-of-the-art in image and video quality assessment with specific emphasis on medical imaging. As described in Chapter 1, quality assessment can be divided into two parts when it comes to an image processing system. On one hand, there is a need to evaluate the quality of image or video affected by noise or other distortions. On the other hand, one must also assess the impact of each processing step on the quality of image. For medical applications like image-guided surgery, these processing steps are image registration, image enhancement and image segmentation. In this chapter, we provide some literature review on both kinds of quality assessment.

When it comes to evaluating the quality of images and videos, human evaluations serve as the ultimate benchmark. However, it is not always convenient or possible to have the human observers available for assessing the quality. To solve this problem, either the model observers or the objective quality metrics are proposed. However, both of these solutions present their own challenges and despite the availability of plenty of work on both of them, there is still a great room for improvement in predicting accurately the image quality. This is especially true in the context of medical imaging. In this chapter, we also present state-of-the-art for all these different aspects such as model observers and medical image quality assessment.

The rest of this chapter is organized as follows. In Section 2.2, we first provide a brief overview on model observers. Thereafter, in Section 2.3, we present a review of the subjective

image quality assessment methods. This is followed by Section 2.4 which describes in detail the state-of-the-art in objective image and video quality assessment metrics. Section 2.5 then gives a literature review on assessment of different image processing algorithms used for image-guided surgeries. After that, in Section 2.6, we describe the benchmarking methods used for evaluating performance of objective metrics. Finally, in Section 2.7, we conclude this chapter.

2.2 Model Observers

Model Observers refer to the mathematical models that are used to predict the task performance of humans. In the context of medical image quality, this task is the detection of a weak signal or tumor in a noisy image [4]. Let us assume that an observed image g is represented in terms of the imaged signal f and additive noise n using the following image signal model [5]:

$$g = \Omega f + n \tag{2.1}$$

where Ω represents an imaging operator. The diagnosis of tumor can then be expressed as a binary classification task having two possible hypothesis:

$$\begin{aligned} \mathcal{H}_0 : g &= \Omega f_b + n \\ \mathcal{H}_1 : g &= \Omega (f_b + f_s) + n \end{aligned} \tag{2.2}$$

where f_b and f_s represent the background and the signal (tumor), respectively. The hypothesis $\mathcal{H}_0$ signifies absence of tumor in this case whereas $\mathcal{H}_1$ implies presence of tumor.

For a model observer, the detection of signal is based on a decision rule that employs a test statistic, $\lambda(g)$ [6]. If this statistic is found to be greater than the criterion λ_c, decision is made in favor of the hypothesis $\mathcal{H}_1$ indicating presence of signal. Otherwise, the hypothesis $\mathcal{H}_0$ for the absence of signal is selected. For evaluating task performance of a model observer, area under curve (AUC) of Receiver Operating Characteristic (ROC) is normally used.

2.2.1 Ideal Observer

The ideal observer (IO) is the one which obtains the highest attainable AUC amongst all model observers. The test statistic for IO is a likelihood ratio test and is given by:

$$\lambda_{IO}(g) = \frac{P(g|\mathcal{H}_1)}{P(g|\mathcal{H}_0)} \tag{2.3}$$

However, IO is only applicable to simple cases, since mostly, for practical applications, it is difficult to compute high-dimensional probability density functions (PDF) needed for the likelihood ratio.

2.2.2 Linear Observer

To avoid non-linear test statistic like likelihood ratio, a linear observer uses a linear discriminant given by:

$$\lambda_{LO}(g) = w^T g \tag{2.4}$$

where w is a real-valued column vector. The Hotelling Observer (HO) provides a practical linear test statistic with

$$w_{HO} = \mathbf{S}_2^{-1}(\langle g|\mathcal{H}_1\rangle - \langle g|\mathcal{H}_0\rangle) \tag{2.5}$$

where $\mathbf{S}_2$ is the intraclass scatter matrix representing average covariance matrix of image:

$$\mathbf{S}_2 = \frac{1}{2}(\mathbf{K}_0 + \mathbf{K}_1) \tag{2.6}$$

In Eq. (2.6), $\mathbf{K}_0$ and $\mathbf{K}_1$ represent ensemble covariance matrices without and with signal respectively. HO is also called the "Ideal Linear Observer" because it gives the maximum value of Signal-to-Noise Ratio amongst all linear observers. Although HO has all the advantages of a linear observer yet the calculation of the inverse $\mathbf{S}_2$ can be problematic even with normal image dimensions.

2.2.3 Channelized Hotelling Observer

To overcome the problem of inverse matrix calculation in HO, Channelized HO (CHO) offers to preprocess image data by decomposing into different channels with the use of a series of filters. Representing linear filtering operation by P filters with $U = [u_1, u_2, ..., u_P]$, the new formalization matrix becomes:

$$g' = U^T g \tag{2.7}$$

and the test statistic is then represented as

$$\lambda_{CHO} = w_{CHO}^T g' \tag{2.8}$$

All the model observers described here work for the task where signal is known exactly. However, for medical images in clinical setting, this is not the case and the task is changed from Signal-Known-Exactly (SKE) to Signal-Known-Statistically (SKS) [6]. This makes the design of model observers for medical image quality assessment a very complex task. Nevertheless, some works have already been done for design of model observers for SKS tasks [7, 8]. We will not go into details of these methods here as in this work, we have tackled the quality assessment problem by proposing new objective metrics rather than by use of model observers.

2.3 Subjective Image/Video Quality Assessment

As mentioned before, the image or video quality can be assessed either subjectively or objectively. In subjective image quality assessment, human observers evaluate the image quality. On the other hand, objective assessment is achieved by estimating perceived image quality automatically using algorithms. Subjective assessment has the drawback of being time consuming and its difficult design [9]. Subjective assessment is also affected by factors like viewing distance, display device, lighting condition, subjects' vision ability, and subjects' mood [10].

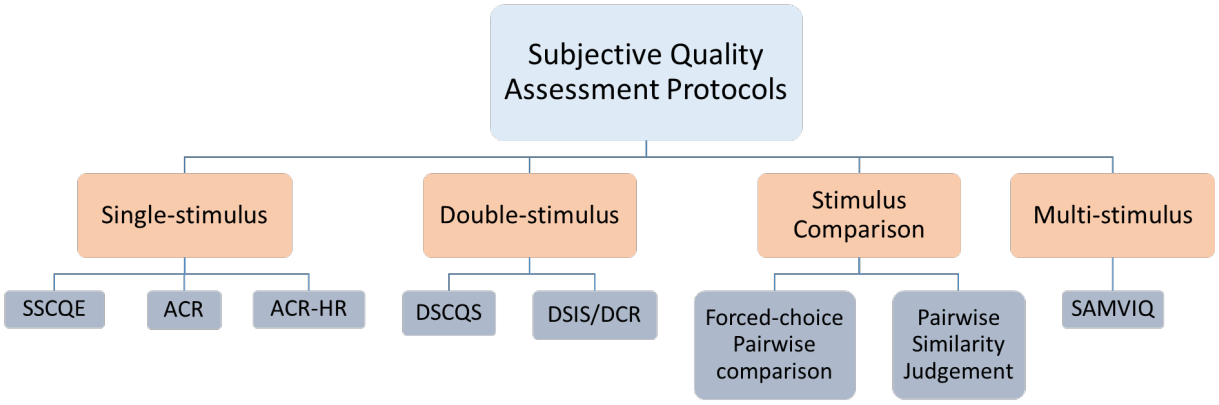


Figure 2.1: Classification of subjective quality assessment protocols

For subjective assessment of image and video quality, different international standards exist with the most common being ITU-R BT.500 [11, 12], ITU-R BT.1788 [13] and ITU-T P910 [14]. All subjective quality assessment works for images and videos rely on using one of these standard

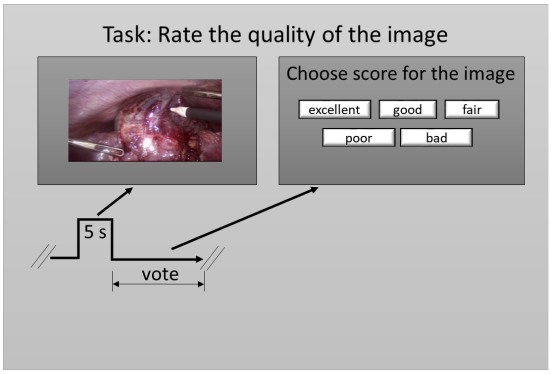
methods [3, 10]. Single and double stimulus methods are the most commonly used amongst them. In general, single and double stimulus methods require rating of test images on a fixed 5-point scale or a continuous scale, in which observers judge the quality of a single and a pair of images [15] respectively. Scores from different observers for each test image are then averaged to get mean opinion score (MOS) and difference mean opinion score (DMOS) after subject and outlier rejection for Single Stimulus and Double Stimulus methodology respectively [3]. To allow an effective comparison of different observers' opinions, calculation of Z-score can also be done for DMOS [10].

2.3.1 Single-Stimulus (SS) Methods

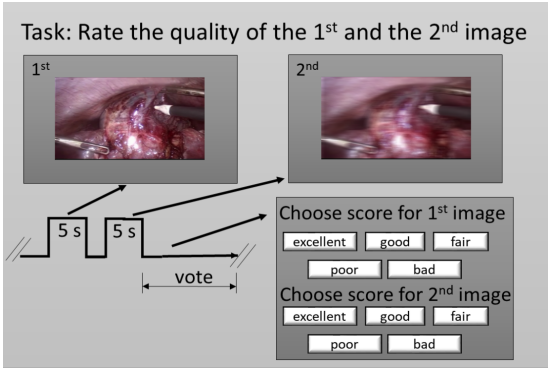
According to the recommendations given by ITU-R BT.500-13 [11], single-stimulus (SS) methods are one of the many methods which can be used for subjective assessment. These are also named alternately as Absolute Category Rating (ACR) by ITU-T P910 recommendations [14]. In single-stimulus method, the assessor is presented with a sequence of test images one after another and he is required to give a score to each image. The sequence of images may or may not contain reference image. If the reference is not known to the observer, the method is called Absolute Category Rating with Hidden Reference (ACR-HR). Usually, each image is shown to the user for a fixed duration after which he is required to give it a rating. The scoring scales can either assess image quality or image impairment. Furthermore, these scales can either be adjectival or numerical. For image quality, five adjectival grades can be Excellent, Good, Fair, Poor and Bad. Five adjectival scales in image impairment scale can be Imperceptible, Perceptible but not annoying, Slightly annoying, Annoying and Very annoying. For image quality, it is also possible to use a continuous scale. However, even in this case it is recommended to also divide the scale into five equal lengths corresponding to five-point quality scale. This method is called Single-stimulus continuous quality evaluation (SSCQE). Figure 2.2a illustrates Single-stimulus method with a discrete adjectival quality scale.

2.3.2 Double-Stimulus (DS) Methods

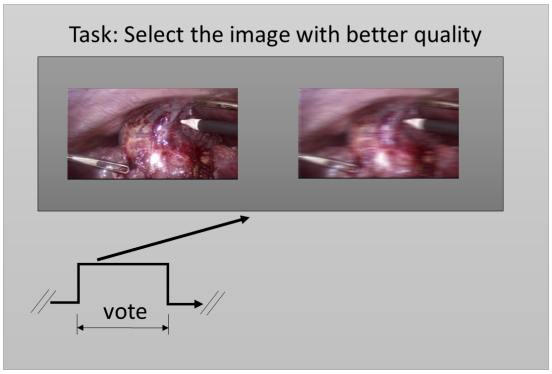
Like single-stimulus methods, the double-stimulus methods fall in one of the two categories depending upon the type of scale used. When the aim is to assess the robustness of systems (i.e. failure characteristics), Double-Stimulus Impairment Scale (DSIS) method is used. This method is also called Degradation Category Rating (DCR) [14]. In this method, the assessor is first presented with an unimpaired image followed by the same image with some impairment added. He is then required to evaluate the second image in comparison to the first one using an



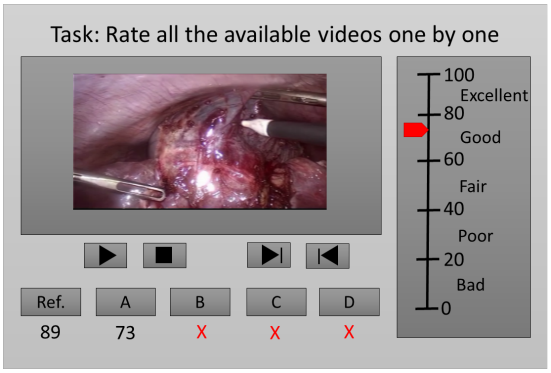
(a) Single-stimulus



(b) Double-stimulus



(c) Pair-comparison



(d) SAMVIQ

Figure 2.2: Subjective quality assessment methods

impairment scale. At the end, mean score for each kind of impairment and for each test image are calculated. This method is suitable for datasets having a complete range of impairments rather than those having a limited range. On the other hand, Double-Stimulus Continuous Quality-Scale (DSCQS) method is the recommended assessment method for applications involving stereoscopic images or where the quality relative to a reference is to be measured. In this method, the assessor is presented with a sequence of image-pairs. In each image-pair, one of the images can be the reference image. However, the assessor is not informed regarding which one is the reference. He is then required to assess the quality of both images using a continuous quality scale. The scores are then normalized in the range between 0 and 100. Wang et al. have shown the application of Double-Stimulus Continuous Quality-Scale (DSCQS) for stereoscopic images in [16], using a sliding adjectival scale. Like for single-stimulus methods, it is also possible to use a discrete adjectival quality scale. Figure 2.2b depicts double-stimulus assessment with such a scale.

2.3.3 Stimulus-Comparison Methods

In these kinds of methods, the assessor is shown a sequence of image-pairs. For each image-pair, he is asked to evaluate the difference between the two images. Like SS method, the scores here can either be adjectival or numerical. Forced-choice Pairwise Comparison (Figure 2.2c) is one such ordering method where observers decide which of the two displayed images has higher quality. Similarly, another method called Pairwise Similarity Judgement requires observers to give difference of quality of two displayed images shown on a continuous scale along with their decision on which one has the higher quality [15].

For pairwise comparison method, subjective scores can either be obtained by averaging all the preferences for an image or a video or by using some sort of psychometric scaling method [17]. The former method can be used when all the pairs in the experiment are compared. Normally, for the latter method, observers are either modeled using Bradley-Terry model [18] or Thurstone's model Case V[19]. In Bradley-Terry model, the difference between the quality of two images/videos is assumed to have a logistic distribution. On the other hand, in Thurstone's model, the rating of each stimulus is assumed to have a Gaussian distribution. Using one of these models, the preference scores are then converted to the scaled scores.

2.3.4 SAMVIQ

SAMVIQ, as shown in Figure 2.2d is a methodology for subjective test of multimedia applications using computer displays proposed in ITU BT 500-14 [12]. It can also be described as a multi stimuli continuous quality scale method with explicit and hidden references. In this method, the

observers can access all the comparison videos of the same category, randomly. They can start and pause the evaluation process whenever they wish, and hence can grade the videos at their own pace. SAMVIQ is also very flexible in the sense that it allows modification of grades for the videos. Moreover, the observers can watch the videos repeatedly if they want.

In SAMVIQ, each distorted video can be compared directly with the reference and hence it provides an absolute measure of the subjective quality. Moreover, with the option to compare each distorted video directly with other impaired versions, all the videos can be graded accordingly. This feature allows a high degree of resolution in the scores given to the systems. Moreover, due to its flexibility, SAMVIQ addresses the problem of possible lack of concentration in methods like DSCQS, where there is a continuous sequential presentation of images or videos. As a result, the errors are reduced and the results are more reliable. However, this flexibility also results in an increase in time for completion of the test.

2.4 Objective Image/Video Quality Assessment

An ideal objective image quality assessment (IQA) method should be able to predict quality as accurate as an average human observer. Depending upon the availability of a perfect reference image having no distortions, the objective IQA algorithms can be divided into three types: (i) full reference image quality assessment (FR-IQA), (ii) reduced reference image quality assessment (RR-IQA) and (iii) no-reference image quality assessment (NR-IQA) [10] [20]. As the names suggest, in full reference IQA methods, the reference image is available, whereas in RR-IQA methods, only a number of features from reference image are available without the full access to such an image. NR-IQA methods do not make use of a reference image. They are blind assessment methods that mainly rely on specific distortions like noise, blur or compression. NR-IQA methods are considered to be ideal for evaluation of medical images [21].

Existing literature related to medical images is limited and mostly focused on MR images due to more recent advancements in MRI technology. Unlike natural images, medical images contain artifacts introduced either due to hardware or human subject. For instance, MR images may contain hardware-related artifacts due to B_0 field inhomogeneity, Gibbs ringing, RF noise, wrap-around, chemical shift, ghosting or electromagnetic interference [21]. Similarly, artifacts in CT include beam hardening, scattering, Zebra, ring, metal artifacts etc. Human-related artifacts may include blurriness due to motion of scanned subjects. An important thing to note in relation to quality assessment of medical images is that a good image quality in terms of a conventional definition, does not necessarily imply a good diagnostic quality. However, an

improved image quality would certainly increase the confidence of a clinician in diagnosis [21]. For medical images, image quality also changes due to variations in acquisition parameters or site location [22]. Dependence on location is due to surrounding conditions like ambience, unintended coupling of receiver coils with any neighboring object and RF transceiver variations.

Objective image quality assessment for medical images requires two different aspects to be considered. One of them concerns the measurement of image quality with respect to artifacts and distortions in the image. The other aspect focuses on evaluation of an image based on its diagnostic value. For instance, there is a possibility that an image is highly distorted yet it has the same or a better diagnostic value than the less distorted image. Following subsections encompass state of the art related to these two aspects.

2.4.1 Conventional Visual Image Quality Assessment

Conventional IQA methods refer to those metrics which predict the perceived quality of an image for a normal observer. Most of the work on conventional IQA has focused on natural images. Here we describe some of the most common and well-known FR and NR IQA metrics.

2.4.1.1 Full-Reference IQA

The two simplest and very common conventional full-reference methods are Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR). MSE (or RMSE), as the name suggests, is the mean (or the square-root of the mean) of the squared differences between intensity values of the test image and the reference image. PSNR is defined as the ratio of the maximum power of a signal and the power of the distortion. However, these two measures are often found to be lacking in conforming to subjective scores [23], since they do not take into account the physiological and psychophysical characteristics of the Human Visual System (HVS).

In order to overcome poor correlation of these metrics, some other FR metrics have been proposed in the literature. Among them, the most common is Structural Similarity Index (SSIM) [23]. SSIM models structural information of an image based on the fact that pixels of a natural image are strongly dependent, with the dependencies carrying useful information about structures of the scene. SSIM is said to achieve its best performance when applied at an appropriate scale that depends on viewing conditions like display resolution and viewing distance. To cater for this, a multi-scale SSIM (MS-SSIM) has also been developed in which image details at different resolutions and viewing conditions are incorporated into the assessment [24]. According to [25], despite its much superior performance as compared to MSE and PSNR, SSIM is not suitable for medical images like CT due to limitations like uniform pooling, distortion underestimation near

hard edges, instabilities in regions of low variance and insensitivity in regions of high intensities. However, studies like [26] [27] and [28] concluded that SSIM performs well for modalities like ultrasound, MRI and X-rays and MS-SIM performs better for capsule endoscopy, for noisy and compressed images.

Another approach used in FR-IQA is called Visual information fidelity (VIF). In this method, natural images from image capturing devices are modeled in the wavelet domain using Gaussian scale mixtures (GSMs). Studies like in [26] and [27] show that VIF is a good evaluation metric for capsule endoscopy and ultrasound for compressed images. Rajagopal et al. [29] have applied four FR-IQA methods, used for assessment of natural images, to the MRI.

Some other FR metrics which have been discussed specifically in relation to the medical images are Shannon's Information Content, Contrast to Noise Ratio (CNR), Quality Index based on Local Variance (QILV), Bhattacharya Coefficient (BC) and Perceptual Difference Models. Shannon's measure as proposed by Fuderer[30] makes use of Shannon's definition of theoretical information content and takes into account Contrast to Noise Ratio (CNR), scan resolution and field of view. It is only dependent on image data and does not take into consideration characteristics of HVS. CNR is also used as a metric for IQA of medical images [21] [31]. It is defined as the ratio between the standard deviation of the pixel values in the Region of Interest to the standard deviation of the noise.

2.4.1.2 No-Reference IQA

Existing No-Reference IQA can be categorized into distortion-specific and general-purpose metrics. Figure 2.3 illustrates broad-level classification of conventional objective IQA methods with special emphasis on NR-IQA. Some of the distortions seen in medical imaging do not exist for natural images. Hence, it makes more sense to highlight a few of the general-purpose metrics here, some of which have also been used for medical IQA.

Many NR metrics use a two-stage framework for quality assessment. In the first step, the distortion type is classified using distorted image statistics and in the second step, the quality score is predicted using the same statistics. The Blind Image Quality Index (BIQI) [32] and the Distortion Identification-based Image Integrity and Verity Evaluation (DIIVINE) index [33] use this two-step framework. BIQI image statistics are extracted from a wavelet transform over three scales and three orientations using Daubechies 9/7 wavelet basis, whereas DIIVINE uses a steerable pyramid decomposition. Next, these feature vectors along with quality scores are used to train individual regression modules (SVR) corresponding to each distortion class. Finally, BIQI or DIIVINE is computed as the product of the probability vector from classifier and the

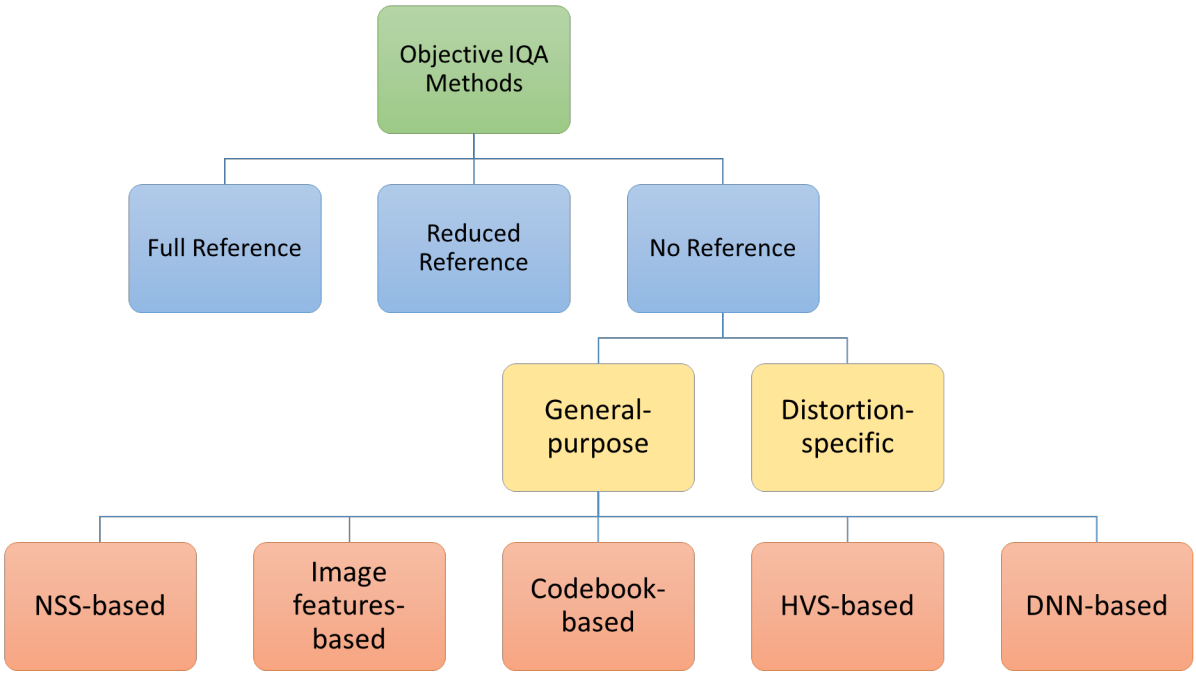


Figure 2.3: Broad-level Classification of objective IQA metrics with special emphasis on NR-IQA

quality vector from regressors.

Another metric called the Blind Image Notator using Discrete cosine transform (DCT) Statistics (BLINDS-II) index [34] uses the Bayesian inference model to predict image quality scores given certain extracted features, which are based on a Natural Scene Statistics (NSS) model of the image DCT coefficients. The calculated NSS features are fed to a regression function to predict the quality. However, all of these methods have limitations like slow computational time due to large numbers of features or statistics. A more competitive and computationally efficient algorithm called Blind/referenceless image spatial quality evaluator (BRISQUE) [35] uses scene statistics of locally normalized luminance coefficients rather than distortion-specific features to quantify possible losses of "naturalness" in the image. Besides these, Natural Image Quality Evaluator (NIQE) [36] is another metric that uses a collection of features extracted from spatial domain NSS model. However, unlike other metrics, NIQE is also opinion-aware and hence does not require subjective scores.

In addition to the use of NSS, in some learning methods like Learning based Blind Image Quality (LBIQ) measure, a regression algorithm is used to incorporate numerous image quality features like natural image statistics, distortion texture statistics and blur/noise statistics. Besides the learning based methods, there are some codebook-based metrics also proposed in literature. For instance, in Codebook Blind Image Quality (CBIQ), Gabor-filter based local features are

extracted from local image patches to capture complex statistics of a natural image, which are then used to calculate DMOS for the entire image. This approach however is computationally expensive and may not accurately represent quality for each patch. Another improved codebook-based method called codebook representation for no-reference image assessment (CORNIA) [37] uses raw image patches as features, resulting in less computation time.

With the rise of deep learning, most recent works are now focused on the use of deep learning for NR-IQA [38, 39, 40, 41, 42]. Besides these, some other NR-IQA methods have also been proposed in literature that are specific to medical images [21]. For instance, Osadebey et al. [43] have proposed a NR metric for quality assessment of brain MRI. Their metric is based on Bayesian framework.

Although there is a plethora of IQA metrics in the literature, yet most of these metrics have been shown to perform well for only limited use-cases. The performance of all these metrics is judged based on some specific quality databases annotated by human observers. However, when the type of images changes or when different distortions are considered, more often than not the performance of these metrics deteriorates. Moreover, most of the databases these metrics are tested with, have synthetic distortions and hence when tested with real distorted images, their performance tends to affect negatively. Also, when it comes to multiple distortions affecting a single image, the performance of most of these metrics decrease sharply. Therefore, it is still a challenge to have a universal quality index that works well in all scenarios. With the limited number of works on medical IQA and with variety of different medical imaging modalities, there is also a great need to propose some new metrics for these types of images and to explore the use of existing metrics on them.

2.4.2 Diagnostic Oriented Image Quality Assessment

Most methods that deal with diagnostic image quality assessment in literature, make use of subjective scores taken from clinicians. These scores are then used as diagnostic quality benchmark to compare with the scores obtained from objective methods applied. For instance, this methodology has been used in works like [26] [27] and [28] which make use of SSIM for modalities like ultrasound, MRI and X-rays and MS-SSIM for capsule endoscopy, for noisy and compressed images.

A very recent study on objective IQA for MR images [44] has shown that SSIM and RMSE perform poorly on these images. In this study, besides these two metrics, the authors have evaluated correlations with radiologists' scores for VIF, MS-SSIM, PSNR, Feature Similarity

(FSIM) index [45] and Noise Quality Metric (NQM) [46] among others. FSIM is a HVS-based FR metric that uses Phase congruency and gradient magnitude as features. NQM, on the other hand, is similar to SSIM in accounting for luminance and contrast differences, but it also considers the effects of contrast masking, spatial frequencies and distance on them. Interestingly, among the metrics tested in [44], the best correlations were found to be of VIF, FSIM and NQM.

In addition, some other methods have also been used for diagnostic IQA that make use of observer models. For instance, Küstner et al [47] have proposed an automated no-reference tool for quality assessment of MRI. This method makes use of a Model observer that gets trained on Human Observer derived labels using Machine learning with deep neural network. In order to reduce the labelling effort of Human Observer, they have proposed an active learning approach which uses a query strategy based on uncertainty sampling.

From the very few works on diagnostic IQA encompassing only some medical imaging modalities, it is evident that there is a big gap to be filled in this field. The difficulty in getting subjective evaluations from medical experts makes this task even more challenging. Furthermore, there is still a lack of clarity on how radiologists and surgeons perceive the quality of medical images as compared to normal observers. This challenge has been taken up recently by carrying out different eye-tracking studies [48]. Such studies can really be helpful in future not only for understanding perception in medical context, but also in developing effective metrics by identifying salient regions in the images.

2.4.3 Stereoscopic Image Quality Assessment

Medical stereoscopic images are the result of stereo endoscopy during laparoscopic surgery. They are essentially two slightly different views of the same image. Stereoscopic IQA methods can be broadly divided into two categories. In the first category, 2D based stereoscopic IQA methods are used for each of the right and left views, without taking into consideration any depth information. The results are eventually combined to give a predicted score. In the second category, disparity information is explicitly used to predict 3D IQA. For stereoscopic IQA of medical images, no-reference methods are the preferred choice because of the lack of availability of reference image.

Campisi et al. [49] were amongst the first ones to propose IQA for stereoscopic images. They used a simple approach of applying 2D FR-IQA to each of the left and right views followed by perceptually inspired approaches to combine the two scores. However, their results suggested that such IQA does not necessarily do well on stereo images.

In another approach, Lin et al. [50] have integrated binocular integration behaviors into existing 2D metrics to propose a new metric for 3D IQA called the Frequency Integrated (FI) metrics. One of the main drawbacks of their proposed method is that the performance of FI-metrics may be affected if there is a conflict between the adopted visual property of a 2D quality metric and the binocular viewing condition. Similarly, Shao et al. [51] have proposed a NR stereo IQA method based on binocular feature combination. Their method uses machine learning and is based on Bayesian theory by assuming that 3D image quality can be modeled by a hybrid combination of posterior and prior features distributions.

Appina et al. [52] in their work have proposed a metric called Stereo Quality Estimator (StereoQUE) which is based on NSS. They have applied a bivariate generalized Gaussian distribution (BGGD) model for the joint distribution of luminance and disparity subband coefficients of natural stereoscopic scenes. For quality score, supervised learning using DMOS scores as labels is performed using support vector machine for regression.

Chen et al. [53] presented a NR IQA algorithm which predicts stereo image quality whether distortion is symmetric or not (rivalrous). Their algorithm utilizes statistical features used for 2D NR algorithms along with binocular rivalry modeled by 3D FR IQA algorithms. The features used are 2D features extracted from Cyclopean image and the 3D features extracted from the estimated disparity map as well as an uncertainty map that is generated by stereo matching algorithm. The quality estimation process follows a two-step procedure similar to the one given in [32].

Hachicha et al. [54] have recently proposed a NR IQA method for stereo image quality assessment. In their method, they apply wavelet transforms using Vector lifting scheme (VLS) [55] to both the views in addition to applying a conventional lifting scheme to the disparity map. Once the wavelet sub-bands are calculated, statistical features are extracted from them using a Generalized Gaussian (GG) distribution approximation. Besides that, further statistical features are also extracted from disparity maps using Bernoulli Generalized Gaussian (BerGG) model. Both these statistical features along with variance and kurtosis are then used to assess the quality of each stereo image using the two-step procedure [32].

While all of these methods have been tested on natural image databases, it would be really interesting to see their performance on medical images. However, unfortunately currently there is no stereo-laparoscopic image quality database available publicly that contains subjective scores. This lack of annotated data is one of the challenges in medical IQA that still needs to be solved.

2.4.4 Video Quality Assessment

There is a prevalent use of laparoscopic and endoscopic videos nowadays, for diagnostics as well as surgical purposes. Assessment of video quality becomes critical in such applications. Moreover, telesurgery [56] is another emerging medical application that needs a good quality video. Like for IQA, most of the existing literature for video quality assessment (VQA) is focused on natural videos. However, some of these metrics developed for natural videos like the Video Multimethod Assessment Fusion (VMAF) metric, have also been employed for medical videos like those from wireless capsule endoscopy and ultrasound [57]. VMAF is a full-reference (FR) metric that does video quality predictions by using compression and scaling artifacts.

For VQA, one simple and common approach is to apply IQA to each frame and then use a temporal pooling strategy like average pooling to get the final quality score of the video. However, extension of the frame-level score to the video score is a challenging task and simple temporal pooling methods may not capture underlying temporal effects accurately. For this reason, temporal pooling strategies based on human psycho-visual reasoning like temporal hysteresis [58] are more promising as compared to simpler ones.

Other kinds of methods for VQA extend the IQA metric for the video. For instance, in [59], Wang et al. have proposed a FR video version of SSIM metric, which measures video quality at three levels namely the local region level, the frame level and the sequence level. Similarly, video BLIINDS (V-BLIINDS) is a NR VQA metric that [60] extends BLIINDS-II to video by using a combination of spatio-temporal NSS-based DCT-domain features and motion coherency features.

Besides these, there are some other metrics that are specific to video like MOtion-based Video Integrity Evaluation (MOVIE) [61] metric and video intrinsic integrity and distortion evaluation oracle (VIIDEO) [36]. MOVIE index is a full-reference metric that tries to capture the characteristics of HVS for video perception by modeling them using separable Gabor filter banks. VIIDEO, on the other hand, is a NR opinion-unaware metric that models statistical naturalness by exploiting space-time statistical regularities. In addition to these, recently many VQA metrics based on deep learning have also been proposed [62, 63, 64, 65, 66] and they have shown promising results for some of the databases on which they have been tested on.

Despite these multiple VQA metrics, none of them works well for all kinds of videos and with all types of distortions, as is the case with IQA metrics. Furthermore, the content of medical videos is starkly different from that in natural videos and it needs to be seen whether some of these VQA metrics also perform well for medical videos.

2.5 Assessment of Selected Image Processing Methods

Limited work has been done when it comes to evaluation of quality of specific processing techniques in a typical image processing pipeline. Following subsections discuss state-of-the-art in each type of assessment associated with image-guided surgery.

2.5.1 Evaluation of Image Registration and Fusion

Use of performance evaluation techniques is also pivotal during different phases in an image processing pipeline such as during image registration. Accuracy of registration process, irrespective of the method used is essential. To measure this accuracy, different errors are proposed in literature namely the localization error, matching error and the alignment error [67]. Localization error occurs due to inaccurate detection of control points (CPs). From simulation studies and ground truth comparisons, mean precision of CP detection methods is well-known for estimating the localization error in a specific case. Matching error is measured by the number of false matches when correspondence between CP candidates is established. Alignment error is referred to as the difference between the mapping models used for the registration and the actual between-image geometric distortion. Different methods exist to estimate alignment error as given in [67].

Generally, the assessment of a fused image can be done either with comparison to a ground truth or without it. Due to lack of ground truth availability, blind assessment is more feasible for medical image fusion. Use of SSIM to get an assessment metric for fusion has been suggested in [68]. Besides structural similarity, spatial similarity between the input images and the fused image has also been used in the performance measure given in [69]. In [70], authors have used Universal Image Quality Index (UIQI) to define three blind fusion metrics. The first metric uses local saliencies to determine the relative importance of one image over another in a window. The second metric is a variant of the first and incorporates overall saliency of a window in accordance with HVS principles. Finally, the last metric further adds edge information to the second metric.

In [71], a blind metric based on mutual information has been proposed for fusion assessment. The metric is evaluated by finding Mutual information (MI) between the fused image and each of the input images, followed by a summation of these MI. To reduce any bias towards the source image with the highest entropy, Hossny et al. [72] have added normalization of the terms before summation.

Edge based assessment methods evaluate the amount of edge information transferred from input images to the fused one. In [73], Sobel edge operator has been used to get the edge strength and orientation from the images. Eventually a relative strength and orientation between input

image and fused image are evaluated to get edge preservation information, the weighted average of which are used to give the performance metric. Similarly in other methods, the gradient information is exploited to get the metric along different directions using spatial frequency [74] and on different scales using Haar wavelet [75].

Besides these, some other metrics for evaluation of fusion inspired from human perception have also been proposed. Among them, Chen-Varshney metric [76] uses edge information, local region saliency and local similarity to get a global quality measure eventually. Another metric, proposed by Chen et al. [77], applies contrast sensitivity filtering in frequency domain followed by local contrast computation. After that, they evaluate the masked contrast map as a measure of contrast preservation. The latter are then used to first compute saliency map and information preservation values for the input images, and then deduce the final score.

2.5.2 Image Enhancement Evaluation

Image enhancement is one of the most important steps, that is required in majority of the image processing tasks. The evaluation of the effectiveness of an enhancement algorithm is challenging, as there is no universal criteria on how to define optimal enhancement. Furthermore, the enhancement process may produce unpredictable side-effects. However, despite these challenges, only few effective metrics have been developed to address this issue of image enhancement evaluation. Several of these works use the same metrics, as those used for conventional IQA, for this task as well. In the context of medical imaging, it is possible to divide image enhancement into three broad categories namely contrast enhancement, denoising and deblurring. Additionally, we also include a fourth category of other enhancement methods which encompasses enhancements like desmoking, dehazing etc. Below we describe the state-of-the-art in evaluation of each of these categories.

2.5.2.1 Contrast Enhancement Evaluation

Medical images like CT, MRI and Ultrasound often require some kind of improvement in contrast in order to better visualize important features like organs and lesions. For assessment of the quality of these enhancement methods, there is a need of contrast enhancement evaluation (CEE) metrics. We can categorize CEE in existing literature into five classes (Figure 2.4) by expanding on the initial categorization done in [78]. Based on this classification, CEE may be statistics-based gradient-based, HVS-inspired, over-enhancement based or fusion-based.

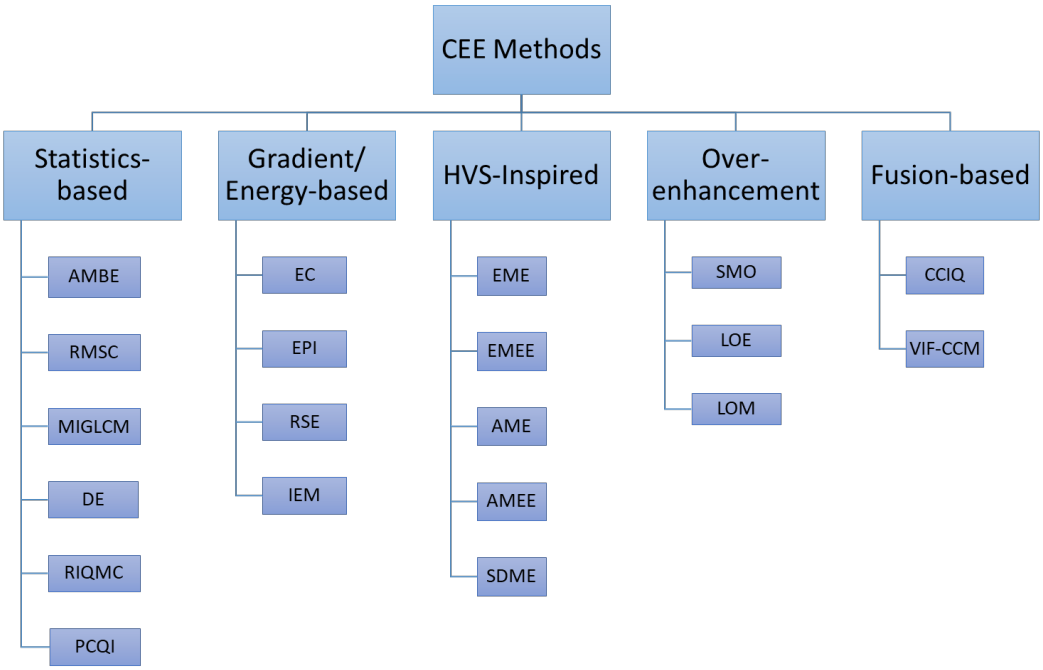


Figure 2.4: Classification of Evaluation metrics for contrast enhancement evaluation

Statistics-Based CEE Metrics

Statistics-based CEE metrics evaluate CE based on the use of statistical properties. Absolute Mean Brightness Error (AMBE) [79] is one such metric, which measures preservation of brightness during enhancement process. It is computed by evaluating the absolute difference in mean intensities of the original and the enhanced images. Besides that, a measure called discrete entropy (DE), measures the image content [80] with a higher value indicative of a richer image. However, both DE and AMBE rely on global image information and fail to take into account the local details and spatial correlation between pixels.

Root mean square contrast (RMSC) [81] is another statistics based CEE. It measures the square root of the average squared differences between each pixel value and the mean intensity. However, this is an over-simplistic way of comparing contrast, since it is possible to get good RMSC value even for an enhanced image with artifacts.

Besides that, Qureshi et al. [82] have proposed another metric called MIGLCM. This metric evaluates the changes in statistical features such as joint entropy and mutual information that are acquired from the grey-level co-occurrence matrix (GLCM) of the original and enhanced images.

Another statistics-based metric by Gu et al. [83], called Reduced-reference Image Quality Metric for Contrast change (RIQMC), is based on two-stage framework. In the first step,

similarity of the two images is compared. They have used Phase Congruency to find important regions before estimating difference in entropy between two images on selected regions. In the second stage, they evaluate first to fourth order statistics to model human comfort. Finally, the overall score is calculated by a simple weighted linear combination.

Besides that, a CEE metric for contrast changed images called the patch-based contrast quality index (PCQI) has been proposed by Wang et al. [84]. In this method, each local patch on the image is decomposed into three independent components namely mean intensity, signal strength and signal structure, so that the distortions can be measured separately. The comparison scores for these three values for the test image with reference image are then combined to give PCQI.

Gradient-Based CEE Metrics

These types of measures rely on information about local signal or energy changes for CEE, since changes in contrast affect the gradient and spectral energy distribution. One such metric is called Edge Content (EC) [85, 80]. It is based on the computation of local gradient of image intensity using the Sobel operator. Another such metric called Edge Preservation Index (EPI) [86] uses Laplacian kernel to evaluate the image gradient.

Besides these, a metric called Radial Spectral Energy [87] works in frequency domain and detects variations in radial spectrum that result from contrast enhancement. Another metric in this category, called Image Enhancement Measure (IEM) [88] first divides the image into overlapping blocks. It is then evaluated as the ratio of sum of absolute differences between center pixel and its eight neighboring pixels in all the blocks of the enhanced image and the corresponding blocks in the original image. However, one major drawback of using gradient-based CEE metrics is that they only consider local changes in the image and do not take into account other important global aspects like brightness preservation.

HVS-Inspired CEE Metrics

These CEE metrics are principally inspired from Weber-Fechner and Michelson contrast measures. One such metric is called the measure of enhancement (EME) [89]. It estimates an average contrast of image by averaging measures based on maximum and minimum intensity values over non-overlapping blocks. A variant of EME uses the measurement of entropy in local contrast and is called measure of enhancement by entropy (EMEE) [89].

Besides these two metrics, Absolute Measure of Enhancement (AME) [89] uses logarithmic Michelson contrast at block-level to measure quality of CE. On the other hand, Absolute Measure

of Enhancement by Entropy (AMEE) [89] utilises entropy in local Michelson contrast. Finally, an HVS-inspired CEE measure called Second Derivative like MEasurement (SDME) metric uses pseudo second-order derivative to evaluate the changes in contrast.

Over-Enhancement Metrics

Evaluation of contrast enhancement is also important in order to detect any issues of over-enhancement. To handle this, some over-enhancement measures have been proposed in literature like the Structure Measure Operator (SMO) [90]. It is insensitive to contrast changes but not to structural changes. SMO is characterized by three components namely edge value, entropy and standard deviation. Non-homogeneity of a pixel is calculated as a product of these three components. SMO is then evaluated as the relative structure change of enhanced image with respect to the original one.

Besides SMO, two other metrics called Lightness Order Error (LOE) [91] and Lightness Order Measure (LOM) [92] detect over-enhancement by evaluating the preservation of lightness order in the image. LOM is the more recent metric that tries to overcome the drawbacks of LOE like that of content dependency.

Fusion-Based CEE Metrics

Fusion-based CEE metrics combine multiple metrics using some form of fusion scheme. For instance, Shokrollahi et al. [93] have proposed in their work a new metric for contrast enhancement evaluation called the contrast-changed image quality (CCIQ). Their metric combine quantities like edge-based contrast criterion (ECC), entropy, correlation coefficient and mean intensity measures using Particle Swarm Optimization (PSO) algortihm for optimization.

In [94], the authors have proposed and used different metrics to evaluate contrast enhancement of mammographic images. First is the Distribution Separation Measure (DSM) which determines the quality of enhancement by using the decision boundary between the target and the background regions. In their second and third proposed metrics named Target-to-Background Contrast Enhancement Measure, they evaluate the change in homogeneity of the mass. The second and third metrics are respectively based on standard deviation and entropy. Finally, they have combined these three metrics to give one single metric.

Recently, in [95] a new CEE metric called VIF-CCM has been introduced. In this method, four metrics of VIF, normalized relative local entropy, linear correlation coefficient and AMBE are fused together using harmonic mean. Although fusion-based approaches seem promising, yet the choice of metrics and the selection of an ideal fusion scheme are challenging tasks.

Besides fusion-based approaches, the authors in [96] have proposed a learning-based approach for enhancement. More specifically, a set of enhanced images are first subjectively scored. They are then used as a training set for learning and applying on other images. However, for robustness of any such learning-based approach, a large number of images with subjective scores are required, which is a non-trivial task.

2.5.2.2 Denoising Evaluation

Noise is one of the most common distortions that can affect a medical image. Moreover, it can appear in many forms like Gaussian noise, Rician noise, Poisson noise, impulsive noise, speckle noise and quantization noise. For denoising, MSE and PSNR [97] have most commonly been used for assessment in the past. Even in a recent review of denoising methods for medical images [98], PSNR is used for evaluation along with SSIM, EPI and FSIM. However, some other denoising metrics can now be found in literature, even those applied specifically to medical images. For instance, Coupe et al. have proposed a metric specific to ultrasound images called the ultrasound despeckling assessment index [99].

For denoising, Buades et al. [100] have made use of three criteria for assessment. One of them is MSE whereas the second one is called the method noise comparison. The method noise depicts which geometrical details are preserved and which are eliminated by a specific denoising algorithm. Finally, the third criterion is the visual quality comparison.

Due to poor correlation of MSE and PSNR with subjective evaluations in general, some other metrics like NQM are also used for assessing performance of denoising methods. Besides that, some variations to PSNR and MSE have also been proposed in the form of Weighted Signal to Noise ratio (WSNR) [46], weighted MSE [101], Visual SNR (VSNR) [102], PSNR-HVS [103] and PSNR-HVSM [104]. As the name suggests, WSNR is the ratio of average weighted signal power to average weighted noise power. VSNR, on the other hand is a two-step metric based on wavelet transform that uses both mid-level and low-level characteristics of human vision. Similarly, PSNR-HVS is an HVS-based variant of PSNR that removes mean shift and contrast-stretching using scanning window, before evaluation of PSNR. The other variant, PSNR-HVSM is similar to PSNR-HVS but uses an improved masking model. Besides these, many more denoising evaluation metrics have been proposed in literature over the last few years. For instance, recently, Lu [105] has proposed a metric based on use of Random forest regression on denoising quality feature vector, that includes features like gradient histogram preservation, variational denoising residual and local self-similarity.

To sum up, a lot of work has been done on development of metrics for denoising evaluation.

However, the focus of most of these works is on noise often seen in natural images like Gaussian noise. For medical images, where there are different other kinds of noise, it would be interesting to evaluate the performance of these metrics and to propose some new and more effective ones.

2.5.2.3 Deblurring Evaluation

Blur is another common distortion present in medical images. In MRI and CT, this can result from motion of patients. In laparoscopic video, it can either be due to out of focus lens or due to fast moving camera. Removal of blur for medical images has been investigated in different works. However, here the focus is only on the evaluation methods for deblurring techniques, as we summarize some of the most common ones below.

The effectiveness of blur removal in images can be evaluated using one of the many deblurring evaluation metrics available in literature. Perceptual Blur Index (PBI) [87] is one such metric that is based on exploitation of limitations of the HVS in blur detection. Similarly, in [106], the authors have introduced the idea of Just Noticeable Blur (JNB) based on human blur perception and have integrated it into a probability summation model for use as a metric. As an extension of this, another metric called cumulative probability of blur detection (CPBD) [107], combines the probabilities of detecting blur at each image by computing a cumulative probability.

For detecting performance of deblurring both locally and globally, a metric called Fast Image SHarpness (FISH) [108] uses the weighted average of the log energies of Discrete Wavelet Transform subbands as an index. Another blur metric, SPArse Representation based Image SHarpness (SPARISH) [109] uses the concepts of sparse image representation and dictionary learning for assessment. Some other blur metrics exploit the effects of blur on image features. For instance, Hassen et al. [110] use the effect of blur on local phase coherence (LPC) structures to evaluate the quality. Similarly, Sang et al. [111] have proposed the metric based on the shape of singular value curve (SVC), which is attenuated due to blur. Another metric called Robust Image Sharpness Evaluation (RISE) makes use of eleven sharpness-aware in spectral and multi-scale spatial features. Besides these, there are some metrics that specifically focus on motion blur like the one proposed by Liu et al. [112]. More recently, deep learning methodologies have also been employed for evaluation of deblurring [113, 114, 115].

Like with other IQA applications, most of the works on medical image deblurring also apply some basic IQA metrics like MSE, PSNR, SSIM and AMBE for performance evaluation of deblurring. However, there have been some limited work on development of blur metrics for medical images. For instance, Lin et al. [116] have proposed a blur metric based on local histogram, called weighted edge width (WEW) for retinal images. Similarly, in [117] a new blur

metric for MRI, called Quadratic Index of fuzzines (QIF) has been proposed and has been shown to be superior to some other natural image based metrics like Maximum Local Variation (MLV) metric. Despite these few available works, there is still a great need to evaluate the effectiveness of current metrics and to develop some new if required, for different medical imaging modalities.

2.5.2.4 Evaluation of Other Enhancement Methods

Besides improvement of contrast and removal of noise and blur, medical images may often require other kinds of enhancement. For example, laparoscopic and dermoscopy images are sometimes affected by uneven illumination. Similarly, endoscopic and laparoscopic images also experience specular reflections, which need to be removed for better visualization [118, 119]. Finally, smoke can also affect the image quality during laparoscopy and hence desmoking algorithms are required to remove smoke from image.

For evaluating illumination enhancement, as is the case in other applications, many works use the basic PSNR and SSIM. However, some metrics have been proposed in literature for evaluating illumination enhancement methods like the one in [120]. Their metric is based on Kullbeck-Leibler Divergence between an estimated Gaussian distribution and the desired one for a given image.

Most existing desmoking algorithms developed for laparoscopic images have been evaluated using MSE, PSNR and SSIM [121, 122, 123]. In [124], they have also used BLIINDS-II and Spatial Spectral Entropy-based Quality (SSEQ) metric for assessment of dehazing in natural images. Similarly, Qu et al. [125] have applied a metric called Perceptual Index (PI) to evaluate their method.

2.5.3 Image Segmentation Evaluation

Evaluation of image segmentation methods is very important to determine the quality of segmentation. These methods can be divided into categories depending upon the availability or non-availability of ground-truth. Supervised evaluation metrics make use of the ground-truth in assessing the quality whereas unsupervised metrics try to determine quality based on other criteria. Figure 2.5 provides a detailed classification of segmentation evaluation metrics based on existing literature.

2.5.3.1 Supervised Evaluation

For image segmentation, assessment methods as proposed in [126] estimate what they call as the goodness of segmentation. Goodness is defined in terms of how well a segmentation method extracts a set of objects. These proposed measures are applicable only in case of supervised

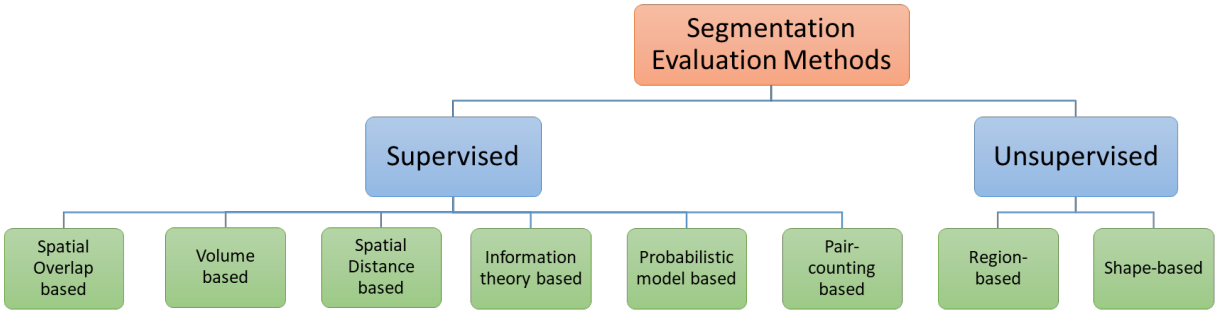


Figure 2.5: Classification of Evaluation metrics for segmentation evaluation

setting. Such measures may be based on methods like polygon matching and often make use of concepts like difference in area and position between reference and segmented objects. These differences can then be used to get a combined metric.

In [127], different metrics have been applied to assess the segmentation performance for tumours in CTs. These metrics include Volume Overlapped (VO), Volume Difference (VD), Average Symmetric Surface Distance (ASD), Root Mean Square Symmetric Surface Distance (RMSD) and Maximum Surface Distance (MSD).

Besides these, some other commonly used metrics for supervised evaluation of segmentation techniques measure similarity based on region overlap. For instance, Jaccard index [128] is such a similarity index which is defined as the ratio of intersection and union of two sets of voxels of same class in two different volumes. Similarly, Dice [129] and Volume Similarity coefficients are other measures of similarity that are based on region overlap.

In [130], 20 different metrics commonly used for evaluation of 3D medical image segmentation have been highlighted. The authors have divided these metrics into six broad categories of measures. These include spatial region overlap based, volume based, pair-counting based, information theoretic based, probabilistic and spatial distance based measures.

The spatial overlap based metrics make use of four basic cardinalities of the confusion matrix namely True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). True Positive Rate (TPR) measure also named as Sensitivity or Recall evaluates the portion of positive voxels in the ground truth which are also identified as positive in test segmentation. In the same way, True Negative Rate (TNR) also called Specificity measures the portion of negative

voxels in ground truth also identified as negative in the test image. Using False Positives and False Negatives instead of True values, result in two other metrics called False Positive Rate (FPR) or Fallout and False Negative Rate (FNR) respectively. Another related metric called Precision or Positive Predictive Value (PPV) is evaluated as the ratio of TP to the sum of TP and FP. Their use for medical images is shown in [131] [132].

The Mutual Information (MI) is an information theoretic based similarity metric that gives a measure of the amount of information one segmented image has about the other. It is evaluated using marginal entropy and the joint entropy. This MI along with entropy can be used to get another metric called the Variation of Information (VOI). It essentially gives information regarding loss (or gain) of information when comparing one clustering partition with another [130].

Among the probabilistic metrics are the InterClass Correlation (ICC), Cohen Kappa Coefficient (KAP) [129], Probabilistic Distance (PBD) [133] and Area under ROC curve (AUC) where ROC stands for Receiver Operating Characteristic. ICC uses correlations of segmentations to give a measure of consistency between two segmentations. KAP on the other hand is a measure of agreement between two segmentations. PBD is often used for fuzzy segmentations and makes use of individual and joint probability distributions for the two segmentations in its evaluation. Finally, AUC is calculated as the area under the ROC curve, which is a plot of True Positive Rate (TPR) to False Positive Rate (FPR).

Lastly, the spatial distance based metrics are dissimilarity measures often used when overall segmentation accuracy is important. The metrics included in this category are the Hausdorff distance (HD) [133], the Average distance (AVD) [134] and the Mahalanobis distance (MHD) [128]. Hausdorff distance is given by the maximum of directed Hausdorff distances between two finite point sets of segmentations. The Average distance is calculated by taking the average of HD over all points. Mahalanobis distance between two points, on the other hand, takes into account correlation of all points in a point cloud.

2.5.3.2 Unsupervised Evaluation

In unsupervised evaluation methods, there is no ground truth. They are based on some criteria needed for a good segmentation. For instance, some of the metrics use intra-region uniformity as a criteria and are based on one of the four quantities of color error, squared color error, texture and entropy for its evaluation [135].

Another category of unsupervised metrics use inter-region disparity between different segments [135]. They make use of quantities like average color difference between regions, local color

difference along boundaries, barycenter distance and layout entropy.

Some unsupervised methods make use of shape information to evaluate quality of segmentation if applicable. One of these methods uses geometrical properties like compactness, elongation and circularity to evaluate regularity of shape.

However, generally the metrics defined for each segment are not used individually but rather in a composite manner. For instance, one way of combining the metrics is to get an unweighted sum of individual per-region metrics. Another method uses a weighted sum with weights proportional to the size of the region. One metric computes weights based on HVS using the human contrast sensitivity curve in Performance Vector (PV) metric [135].

2.6 Benchmarking performance of IQA

Many performance indicators are used to evaluate the performance of the predicted scores from a proposed metric in comparison to subjective scores. The most common among them are Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC), Kendall Rank-Order Correlation Coefficient (KROCC), Outlier Ratio (OR), Mean Absolute prediction Error (MAE) and Root Mean Square prediction Error (RMSE).

OR, MAE and RMSE show the prediction consistency of a given IQA method, whereas PLCC, SROCC and KROCC reflect its good correlation with human judgments. Thus, lower values of OR, MAE, RMSE and higher PLCC, SROCC and KROCC values (closer to 1) indicate good performance [54].

2.6.1 Pearson Linear Correlation Coefficient (PLCC)

PLCC is used to calculate prediction accuracy between two sets of data, with one of them being the reference data. In general terms, it is a measure of linear correlation between two variables and its value lies between -1 and 1. A negative value represents a negative correlation, 0 represents no correlation whereas a positive value indicates a positive correlation. In context of IQA algorithms, the predicted scores can be compared to subjective scores using PLCC. Given that the dataset of subjective scores is represented as $\{x_1, x_2, ..., x_n\}$ and the predicted score dataset is depicted as $\{y_1, y_2, ..., y_n\}$, then r the PLCC can be calculated using the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.9}$$

where $\bar{x}$ and $\bar{y}$ are the means of subjective and predicted scores respectively and n is the number of values in each dataset.

2.6.2 Spearman Rank-Order Correlation Coefficient (SROCC)

SROCC can be used to measure the prediction monotonicity of an IQA metric. In order to calculate SROCC, both the original and the predicted datasets are first sorted and assigned ranked values ranging from 1 till n, where n is the number of values in the datasets. Pearson correlation coefficient for these new datasets is then calculated to give SROCC, r_s. A simplified equation commonly used for calculation of SROCC results in cases where all n ranks are distinct integers. For this, first the difference, d_i between two ranks for each value i (corresponding to each image) is calculated. The equation for SROCC in simplified form is given below:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{2.10}$$

The value of SROCC also lies between -1 and 1. A negative value signifies a decreasing monotonic trend between the two datasets, whereas a positive value implies an increasing monotonic trend. A decreasing trend means that the one variable is inversely proportional to the other. On the other hand, an increasing trend implies that both variables are proportional. A zero value signifies that the two variables are independent.

2.6.3 Kendall Rank-Order Correlation Coefficient (KROCC)

KROCC, like SROCC is also a measure of monotonicity between the predicted metrics and the subjective values. It uses the concept of concordant and discordant pairs. Given two sets of subjective scores and predicted scores represented as $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$ respectively, a set of observations for the two sets can be depicted as (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). A pair of observations (x_i, y_i) and (x_j, y_j) from this set is said to be concordant if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. On the other hand, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$, the pair is said to be a discordant pair. A pair is neither concordant nor discordant if $x_i = x_j$ or $y_i = y_j$. Using the number of concordant pairs in a dataset, n_c and the number of discordant pairs, n_d, KROCC, τ is defined as:

$$\tau = \frac{2(n_c - n_d)}{n(n-1)} \tag{2.11}$$

2.6.4 Outlier Ratio (OR)

This is simply defined as the ratio of 'false' scores predicted by an IQA metric to the total number of scores. A 'false' score is generally taken as a score that lies outside the interval

$[score_S - 2\sigma, score_S + 2\sigma]$, where $score_S$ is the subjective score and σ is the standard deviation of the subjective score. A low value of OR signifies a better performance of a specific IQA method.

2.6.5 Mean Absolute prediction Error (MAE)

MAE is simply a measure of the difference between the predicted score and the subjective score. A smaller value of MAE signifies a better performance of an algorithm. Given that the dataset of n subjective scores is represented as $\{x_1, x_2, ..., x_n\}$ and the predicted score dataset is depicted as $\{y_1, y_2, ..., y_n\}$, MAE is calculated as:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{2.12}$$

2.6.6 Root Mean Square prediction Error (RMSE)

RMSE, as the name suggests is also a measure of difference between the predicted and the subjective score. Like MAE, a smaller RMSE value is indicative of a better performance by an IQA algorithm. Given that the set of subjective scores is represented as $\{x_1,.....,x_n\}$ and the predicted score set is depicted as $\{y_1,.....,y_n\}$, RMSE is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{2.13}$$

2.7 Conclusion

In this chapter, we have reviewed the main issues in image and video quality assessment in the context of natural images as well as medical ones. More precisely, a state-of-the-art of the subjective and objective quality assessment methods has been addressed. We have also reminded the standard metrics used to evaluate the performance of the developed quality assessment algorithms. Based on the discussion in this chapter, it can be safely concluded that the quality assessment aspect for medical imaging has still not been as extensively addressed as many other aspects related to it like segmentation, especially in the context of different imaging modalities involved. While this review has been made for different image modalities, we will focus in the next chapter on the quality assessment of the stereo image modality.

CHAPTER 3

Joint Statistics Based Stereo Image Quality Assessment

Abstract

The recent advances in 3D acquisition and display technologies have led to the use of stereoscopy for a wide range of applications including laparoscopic surgery. The quality assessment of such stereo data becomes of great interest especially when the reference image is not available. For this reason, we present in this chapter a novel no-reference 3D image quality assessment algorithm based on joint statistical modeling of the wavelet subband coefficients of the stereo pairs. More precisely, we resort to bivariate and multivariate statistical modeling of the texture images to build efficient statistical features. These features are then combined with the depth ones and used to predict the quality score based on machine learning tools. Due to non-availability of labeled medical data for quality assessment, the proposed methods are evaluated on a natural image database called LIVE 3D. The obtained results show the good performance of joint statistical modeling based approaches[1].

[1] [136] **Khan, Z. A.**, Kaaniche, M., Beghdadi, A. and Alaya Cheikh, F., 2018, November. "Joint statistical models for no-reference stereoscopic image quality assessment". In 2018 7th European Workshop on Visual Information Processing (EUVIP) (pp. 1-5). IEEE.

3.1 Introduction

Stereoscopic imaging has a wide set of uses ranging from commercial applications like 3DTV and movies [137] to medical applications like laparoscopic/endoscopic surgery [138]. A stereoscopic image (SI) is a pair of views, referred to as left and right images, captured from two different viewpoints. The main advantage of using SI is its ability to provide the 3D (depth) information of the observed scene, based on the disparity existing between the two views. Due to their ever increasing demand and use for consumer-based and critical applications, the assessment of stereoscopic image quality remains a task of an utmost importance.

Many approaches have been discussed in literature to address the problem of objective quality assessment for stereoscopic images, although the number of such works is far fewer than those for 2D monoscopic images. As with 2D image quality assessment (IQA) methods, stereoscopic image quality assessment (SIQA) also falls into one of the three classes of full-reference (FR), reduced-reference (RR) or no-reference (NR), corresponding respectively to cases where there is availability, partial availability or no availability of a reference image or its related information. However, due to the non-availability of a pristine reference image in many applications, it would be more interesting to design a no-reference SIQA.

SIQA (also referred to as 3D IQA) methods can be broadly divided into two categories. In the first one, 2D based IQA methods are used for each of the right and the left views, without taking into consideration the depth information [139][140]. The results are eventually combined to give a predicted score. In the second category, disparity information is used explicitly to predict the quality of the stereo pairs [141][54].

The use of wavelet transforms for IQA is a very common approach both for 2D as well as 3D images because of their correlation with HVS characteristics. For instance, Zhu *et al.* [142] have formulated a SIQA method based on wavelet decomposition, contrast conversion and masking. Moreover, the recent algorithms developed for NR SIQA aim at extracting statistical features followed by a two-stage framework based on machine learning, similar to the one proposed by Moorthy *et al.* [32]. For instance, methods proposed by Chen *et al.* [53] and Hachicha *et al.* [54] follow the same quality prediction stage and result in much better results, while the one developed in [54] outperforming all other methods. Indeed, in [53], 2D texture features are extracted from a cyclopean image (which represents a binocular combination of the left and right views) and 3D features are extracted from the estimated disparity map as well as an uncertainty map that is generated by a standard stereo matching algorithm. Note that the use of cyclopean image to design 3D IQA has also been investigated in [143] [144]. On the other hand, Hachicha

et al. [54] use texture features from the two views after applying a joint wavelet decomposition based on Vector Lifting Scheme (VLS) [55], as well as depth features extracted from the wavelet representation of the estimated disparity map.

The use of statistics for SIQA has been further investigated in recent years by resorting to bivariate statistical models [145] [52]. For instance, Su *et al.* [145] propose to extract the statistical features of a bivariate generalized Gaussian (BGG) model to exploit the spatial image correlation between adjacent subband coefficients of the cyclopean image. Inspired from this work, Appina *et al.* [52] resorted to the same statistical distribution to model the joint distribution between each image luminance and its associated disparity subband coefficients. However, the correlation between a texture image and its depth information is not high enough due the specific characteristics of these inputs. Thus, higher correlation between other inputs could be found to design more efficient joint statistical features for 3D IQA.

In a more recent work, Sinno et al. [146] have explored modeling of adjacent pixels for bandpass-filtered luminance images using BGG. However, they have developed this method for 2D images only. On the other hand, another very recent work by Yao et al. [147] have exploited bivariate statistics of 3D images for SIQA. Their method involves development of their own specific 3D representation based on disparity map and the two views, followed by classification of different regions of the representation into one of the three defined classes. However, their method exploits local information only and the bivariate modeling they use is based on estimation of marginal and conditional probabilities of bins of neighboring responses.

In this chapter, we propose to use joint statistical models for NR 3D IQA. More precisely, we resort to bivariate as well as multivariate statistical models to *simultaneously* exploit the inter- and intra-image redundancies. To this end, a symmetric joint wavelet decomposition is first applied to the stereo pairs to generate the wavelet representations of the left and right views. Then, statistical features resulted from the joint statistical modelling of both subband coefficients are extracted using the bivariate and multivariate generalized gaussian distribution. Finally, these features are combined with the depth ones and used to predict the quality of the stereo images.

The remainder of this chapter is organized as follows. Section 3.2 presents a related previous work, and Section 3.3 describes the adopted methodology in this work. Section 3.4 illustrates the results of the proposed methods. Finally, some conclusions and future work are drawn in Section 3.5.

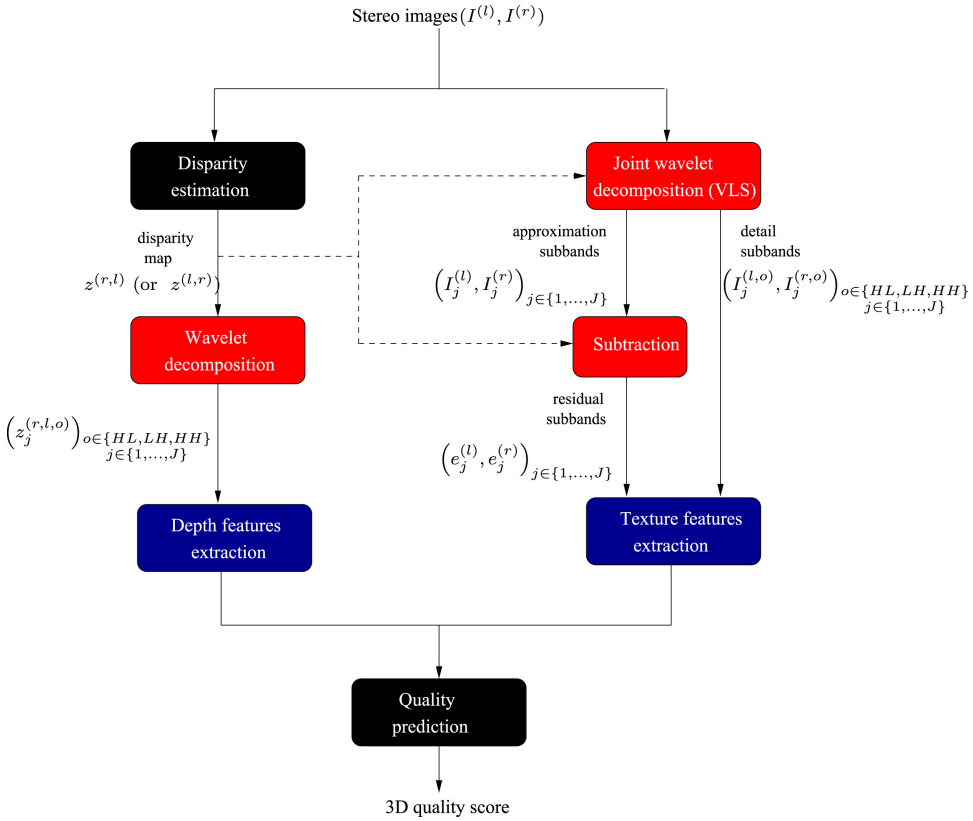


Figure 3.1: Block diagram of a recent NR 3D IQA metric [54].

3.2 Previous Work

Our framework for 3D quality assessment is built upon the work developed in [54]. The flowchart of the latter method is shown in Figure 3.1. In this method, two different kinds of features are used for quality assessment namely the texture features and the depth features. For texture features, a joint wavelet decomposition is first performed for the stereo images using the estimated disparity map. Then, the coefficients of detail subbands and the residual subbands generated from this decomposition are modeled using a suitable probability distribution. The parameters of this distribution are then used as the texture features.

For depth features, first the diparity map is estimated from the stereo images using a Disparity Estimation algorithm (DE). Then, the disparity map is transformed using a wavelet decomposition. Finally, the depth features are extracted from the parameters of the distributon used to model subband coefficients. The main blocks of this method are described in more detail in the following sections.

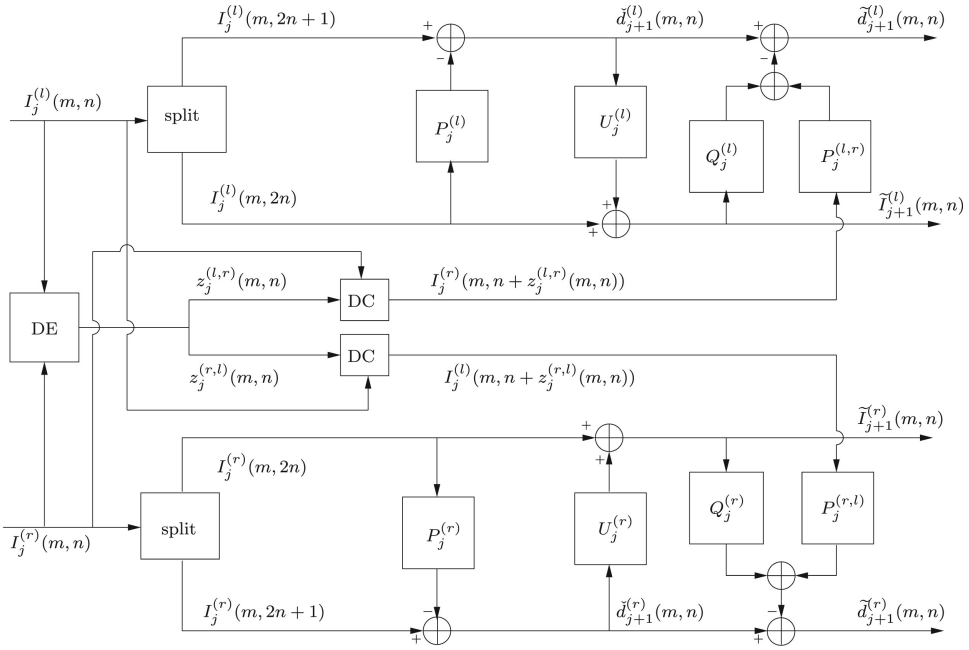


Figure 3.2: VLS joint wavelet decomposition [54].

3.2.1 Joint Wavelet Decomposition

As shown in Figure 3.1, for a given stereo pair $(I^{(l)}, I^{(r)})$, the disparity maps are firstly generated using a standard approach. Then, a symmetric joint wavelet decomposition, based on Vector Lifting Scheme (VLS) [55], is applied to both views. This joint decomposition aims to exploit the cross-view redundancies with the help of the estimated disparity map.

VLS is a 3D wavelet decomposition which has been adapted to the context of stereoscopic data. Figure 3.2 provides a flow-chart of the VLS based joint wavelet decomposition used. For each row m, two Predict-Update-Predict (PUP) lifting structures are applied for the given resolution, j to the right and left images $I_j^{(r)}$, $I_j^{(l)}$. The first prediction step is used to generate an intermediate detail signal at a resolution level $j+1$. For the right image, this detail signal $\check{d}_{j+1}^{(r)}$ is given by

$$\check{d}_{j+1}^{(r)}(m,n) = I_j^{(r)}(m,2n+1) - \sum_{k \in \mathcal{P}_j^{(r)}} p_{j,k}^{(r)} I_j^{(r)}(m,2n-2k) \tag{3.1}$$

where $p_{j,k}^{(r)}$ and set $\mathcal{P}_j^{(r)}$ represent the weights and support of the odd sample predictor for right image, respectively and n represents image column. After this, in the update step, approximation coefficients are computed as

$$\tilde{I}_{j+1}^{(r)}(m,n) = I_j^{(r)}(m,2n) + \sum_{k \in \mathcal{U}_j^{(r)}} u_{j,k}^{(r)} \check{d}_{j+1}^{(r)}(m,n-k) \tag{3.2}$$

where $u_{j,k}^{(r)}$ and set $\mathcal{U}_j^{(r)}$ represent the weights and support of the update operator for right image, respectively. At the end, the second prediction step is performed to get the final detail coefficients, $\tilde{d}_{j+1}^{(r)}$

$$\begin{aligned} \tilde{d}_{j+1}^{(r)}(m,n) = & \check{d}_{j+1}^{(r)}(m,n) - \Bigg(\sum_{k \in \mathcal{Q}_j^{(r)}} q_{j,k}^{(r)} \tilde{I}_{j+1}^{(r)}(m,n-k) \\ & + \sum_{k \in \mathcal{P}_j^{(r,l)}} p_{j,k}^{(r,l)} I_j^{(l)}(m,2n+1+z_j^{(r,l)}(m,2n+1)-k) \Bigg) \end{aligned} \tag{3.3}$$

where $q_{j,k}^{(r)}$ and $\mathcal{Q}_j^{(r)}$ are respectively the weights and support for the second *intra*-image predictor for right view whereas $p_{j,k}^{(r,l)}$ and $\mathcal{P}_j^{(r,l)}$ represent the weights and support for the *inter*-image predictor respectively. In Eq. (3.3), $z_j^{(r,l)}$ is derived from the initial estimated disparity map, $z^{(r,l)}$ by sampling and dividing by 2^j, since the original image dimensions are also divided by 2^j at the j-th resolution level. The superscript (r,l) here represents that the values in the disparity map allows to find for each pixel in the right image its homologous one in the left image. The disparity map, $z_j^{(r,l)}$ is then used in the disparity compensation (DC) process to compute $I_j^{(l)}(m,n+z_j^{(r,l)}(m,n))$ used in the second hybrid prediction stage.

Eventually, this joint decomposition is applied to each row and each column of the image. A similar decomposition is also applied to the second view (i.e. the left one) as shown in Figure 3.2. As a result, we obtain one approximation subband and three detail subbands with horizontal, vertical and diagonal orientations for each image. By applying the same decomposition process on the approximation subbands of both views, we get finally two wavelet representations of the left and right images over J resolution levels.

3.2.2 Texture Feature Extraction

The resulting detail subband coefficients of the left and right views are then modeled using a generalized Gaussian (GG) distribution. In addition to the detail subbands, and since the approximation subband coefficients can not be well modeled by a GG distribution, it has been proposed to generate a residual subband for each view by computing the difference between its original approximation subband and the disparity compensated one of the other view. Then, the GG distribution has been used to model the obtained residual subband coefficients. Let us recall

that the probability density function of GG distribution for the j-th subband with orientation o is given by:

$$\forall \quad \xi \in \mathbb{R}, \quad \forall \quad v \in \{l, r\},$$
$$\tilde{h}_j^{(v,o)}(\xi) = \frac{\beta_j^{(v,o)} (\omega_j^{(v,o)})^{1/\beta_j^{(v,o)}}}{2\Gamma(1/\beta_j^{(v,o)})} e^{-\omega_j^{(v,o)} |\xi|^{\beta_j^{(v,o)}}}$$

where Γ is the gamma function defined by:

$$\forall \quad x > 0, \qquad \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt, \tag{3.4}$$

$\omega_j^{(v,o)}$ and $\beta_j^{(v,o)}$ are respectively the scale and shape parameters for the left view (i.e "v" $= l$) and the right one (i.e "v" $= r$).

The scale and shape parameters are estimated using moment matching method and are combined with variance $\sigma_j^{(v,o)}$ and kurtosis $\kappa_j^{(v,o)}$ of each subband to construct the following texture feature vector:

$$\mathbf{f} = (\mathbf{f}^{(l)}, \mathbf{f}^{(r)}) \tag{3.5}$$

where for each $v \in \{l, r\}$,

$$\mathbf{f}^{(v)} = \left(\beta_j^{(v,o)}, \omega_j^{(v,o)}, \sigma_j^{(v,o)}, \kappa_j^{(v,o)}\right)_{\substack{o \in \{HL,LH,HH,LL\} \\ j \in \{1,\ldots,J\}}}. \tag{3.6}$$

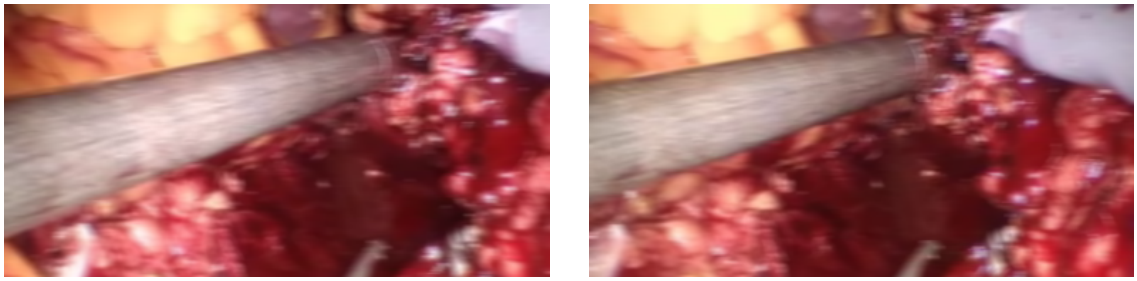
3.2.3 Depth Feature Extraction

Besides texture features, depth ones have been extracted. To this end, both the disparity maps with respect to the left and the right views, are first estimated using the variational method given in [148]. Some examples of disparity maps estimated using this method are shown in Figure 3.4 for images from Figure 3.3. Next, these disparity maps are transformed into the wavelet domain using 5/3 lifting scheme. The detail subbands of this representation are then modeled using a Bernoulli generalized Gaussian (BerGG) model whose probability density function is given by:

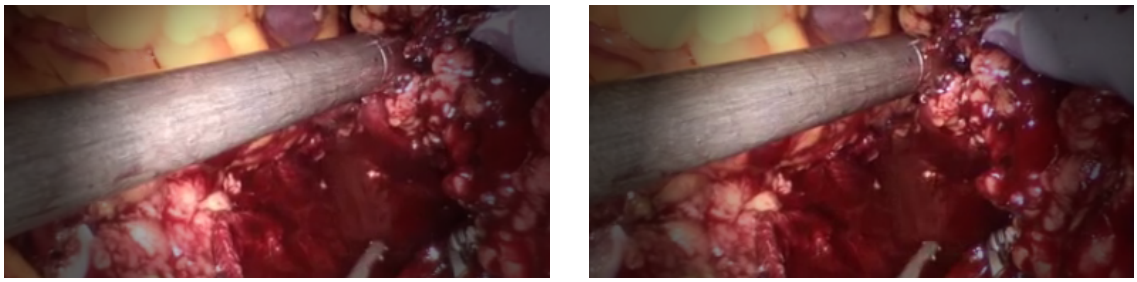
$$\forall \xi \in \mathbb{R}, \qquad h_j^{(z)}(\xi) = (1 - \epsilon_j^{(z)}) \delta(\xi) + \epsilon_j^{(z)} \tilde{h}_j^{(z)}(\xi), \tag{3.7}$$

where $\tilde{h}_j^{(z)}(\xi)$ represents the probability density function of a GG distribution, $\epsilon_j^{(z)} \in [0, 1]$ denotes the mixture parameter and δ is the Dirac distribution. Figure 3.5 shows the diagonal detail coefficients of dispartiy maps in Figure 3.4 modeled by BerGG.

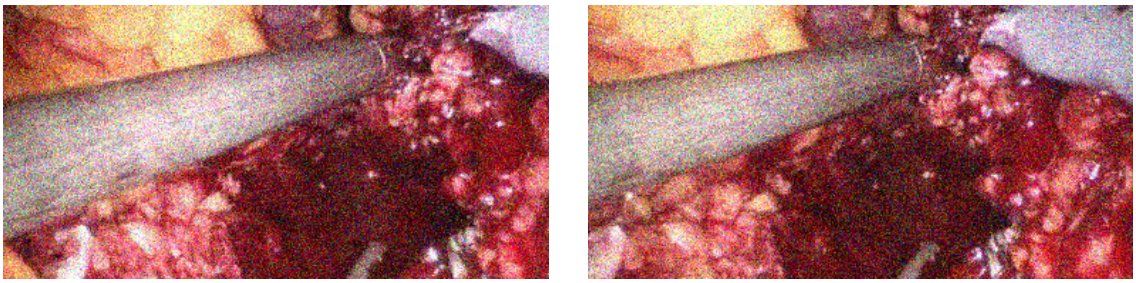
(a) (b)



(c) (d)

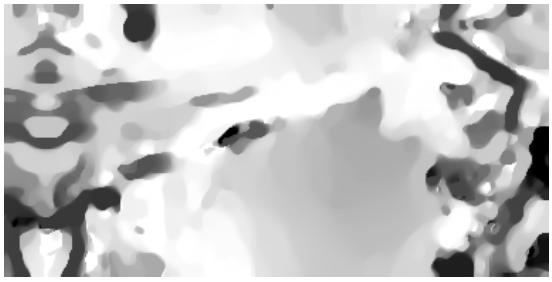


(e) (f)

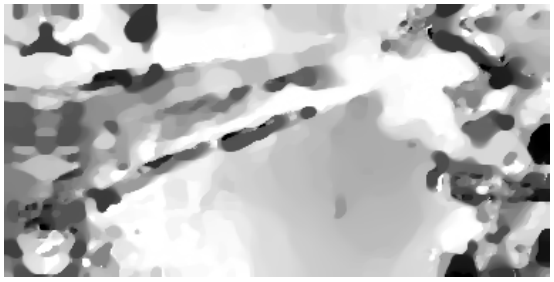


(g) (h)

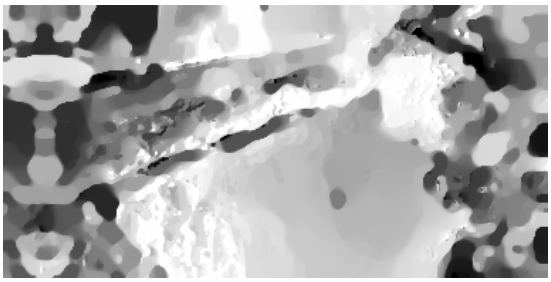
Figure 3.3: Distorted laparoscopic stereo-pairs with distorions (top to bottom): Defocus blur, Motion blur, Uneven illumination and noise



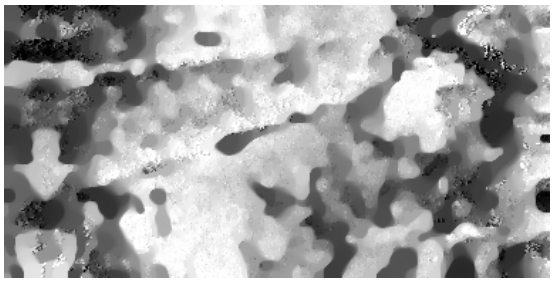
(a) Defocus blur



(b) Motion blur



(c) Uneven illumination



(d) Noise

Figure 3.4: Estimated disparity maps based on variational approach for images in Figure 3.3

Using the moment matching method, the estimated parameters of the BerGG model as well as the variance and kurtosis parameters are then combined to construct the following set of depth features:

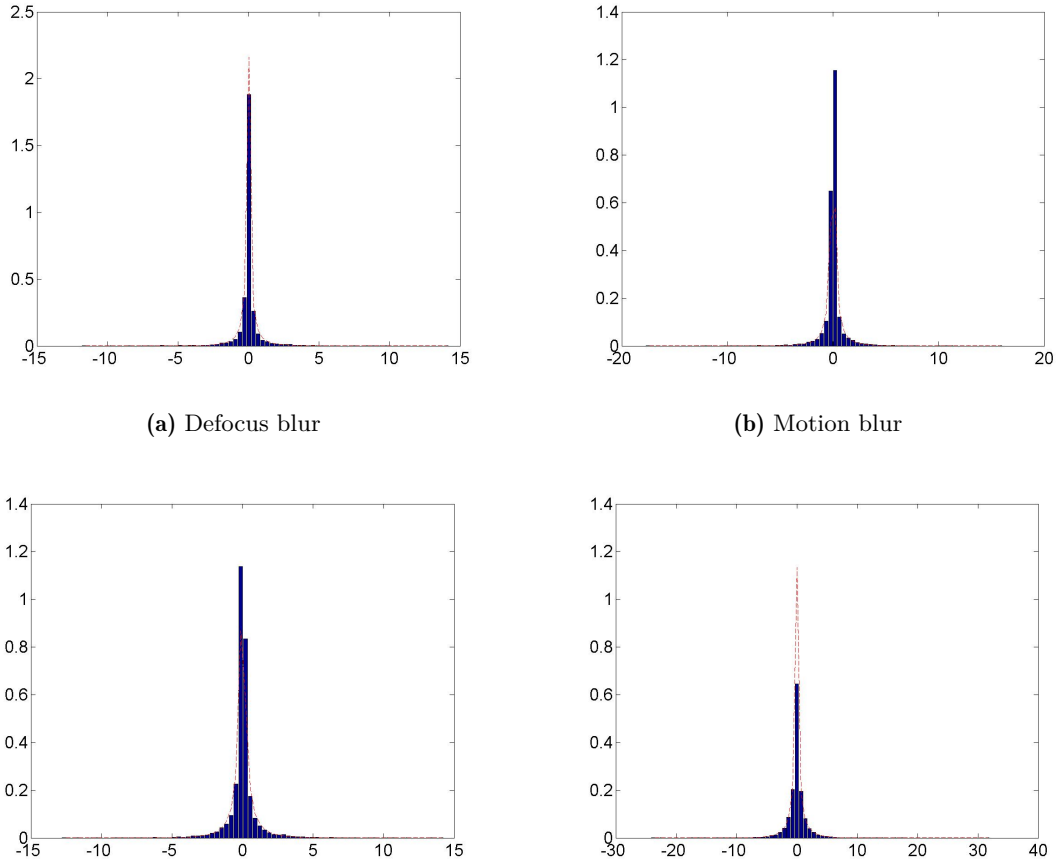
$$\mathbf{f}^{(z)} = \left(\epsilon_j^{(z,o)}, \beta_j^{(z,o)}, \omega_j^{(z,o)}, \sigma_j^{(z,o)}, \kappa_j^{(z,o)}\right)_{\substack{o\in\{HL,LH,HH\} \\ j\in\{1,\ldots,J\}}} \tag{3.8}$$

3.2.4 3D Quality Evaluation

To predict the quality of the stereo images, a two-stage framework based on machine learning tools, similar to the one proposed in [32] [53], is used. More precisely, once the texture and depth features are obtained for a set of distorted images, they are used to train a Support Vector machine (SVM) along with their distortion classes as input. Moreover, for each distortion type, a separate Support Vector Regressor (SVR) is trained using these features and their subjective DMOS as input. Finally, the scores for new test images are obtained by first calculating their features and then passing them through the trained SVM and SVRs.

3.3 Proposed Joint Statistics based SIQA Method

Based on the previous work described, we propose here to utilize the same methodology but with more effective and relevant texture features. As explained later in this section, these features can



(a) Defocus blur

(b) Motion blur

(c) Uneven illumination

(d) Noise

Figure 3.5: Diagonal detail wavelet coefficients of disparity maps of Figure 3.4 modeled by BerGG.

be derived from a joint statistical modeling scheme. In our proposed work, we have utilized two different kinds of joint statistical modeling namely the bivariate and the multivariate modeling. In the following, we describe in more detail the motivation and the details of our proposed method.

3.3.1 Motivation

One can notice that the previous work presented in [54] makes use of texture features extracted separately from the left and right views. However, such *univariate* statistical modeling is not so efficient since it does not exploit the inter-views dependencies. Therefore, we first propose to resort to *bivariate* statistical modeling of the wavelet subbands of both views to construct more efficient features. Then, in order to further take into account the spatial redundancies existing between adjacent wavelet coefficients in each subband, we investigate *multivariate* statistical modeling for SIQA. Therefore, compared to the previous work, it is important to note here that

our main contribution here concerns the texture feature extraction step while the two remaining ones (depth feature extraction and 3D quality evaluation steps) are kept unchanged.

3.3.2 Bivariate statistical modeling based texture feature extraction

In order to exploit the joint statistics existing between the wavelet subbands of the left and right views, we propose to use the bivariate Generalized Gaussian (BGG) distribution. Note that the latter is a particular case of the multivariate GG (MGG) distribution. To introduce this model, let us denote by $\tilde{\mathbf{w}}_j$ a multivariate vector composed of c statistically dependent components. We assume that the set of coefficients vector $\tilde{\mathbf{w}}_j$ constitutes an independent identical distributed sample of a random vector $\tilde{\mathbf{W}}_j$. Thus, the probability density function $h_{\tilde{\mathbf{W}}_j}$ of the MGG distribution is given by [145]:

$$\begin{aligned} &\forall \quad \mathbf{w} \in \mathbb{R}^c, \\ &h_{\tilde{\mathbf{W}}_j}(\mathbf{w}; \mathbf{\Sigma}_j, \alpha_j, \beta_j) = \frac{1}{|\mathbf{\Sigma}_j|^{\frac{1}{2}}} g_{\alpha_j,\beta_j}(\mathbf{w}^\top \mathbf{\Sigma}_j^{-1} \mathbf{w}) \end{aligned} \tag{3.9}$$

where $\mathbf{\Sigma}_j$ is a $c \times c$ symmetric scatter matrix, α_j and β_j are the scale and shape parameters respectively, and $g_{\alpha_j,\beta_j}(.)$ is referred to as the density generator expressed as follows:

$$g_{\alpha_j,\beta_j}(\mathbf{w}^\top \mathbf{\Sigma}_j^{-1} \mathbf{w}) = \frac{\beta_j \Gamma(\frac{c}{2})}{(2^{\frac{1}{\beta_j}} \pi \alpha_j)^{\frac{c}{2}} \Gamma(\frac{c}{2\beta_j})} e^{-\frac{1}{2}(\frac{\mathbf{w}^\top \mathbf{\Sigma}_j^{-1} \mathbf{w}}{\alpha_j})^{\beta_j}} \tag{3.10}$$

In our first joint statistical modeling, the bivariate vector $\tilde{\mathbf{w}}_j^{(r,l,o)}$ is defined as follows:

$$\tilde{\mathbf{w}}_j^{(r,l,o)} = (I_j^{(r,o)}, I_j^{(l,o)})^\top \tag{3.11}$$

where $I_j^{(v,o)}$ is the j-th wavelet subband, with orientation o, of the view $I^{(v)}$.

To increase the correlation between the wavelet subbands of the left and right views, a pre-processing step is performed by removing the occluded areas. This is done by removing the leftmost pixels from the subbands of left view and the rightmost pixels from the subbands of right view. Then, the BGG distribution (particular case of the MGG one, with $c = 2$) is used for statistical modeling and the resulting parameters $\left(\alpha_j^{(B,o)}, \beta_j^{(B,o)}\right)$ are estimated using the moment matching method [149]. Note the the superscript B is used to refer to the statistical parameters extracted in the bivariate modeling case.

Finally, the texture feature vector is built by taking the scalar coefficients of the estimated statistical parameters:

$$\mathbf{f}^{(B)} = \left(\alpha_j^{(B,o)}, \beta_j^{(B,o)}\right)_{\substack{o \in \{HL,LH,HH,LL\} \\ j \in \{1,\ldots,J\}}} \tag{3.12}$$

3.3.3 Multivariate statistical modeling based texture feature extraction

In addition to the cross-view similarities, the spatial redundancies in each wavelet subband could be further exploited. To this end, we propose to resort to a multivariate statistical modeling using the MGG distribution. In this case, our multivariate vector $\tilde{\mathbf{w}}_j^{(r,l,o)}$ will be constructed by taking, for each subband j, the set of the wavelet coefficients of the left image $I_j^{(l,o)}$ and the right one $I_j^{(r,o)}$, located at the same spatial position (m,n), as well as those of the neighboring pixels:

$$\tilde{\mathbf{w}}_j^{(r,l,o)} = \left(I_j^{(r,o)}, I_{j,p,q}^{(r,o)}, I_j^{(l,o)}, I_{j,p,q}^{(l,o)}\right)^\top \tag{3.13}$$

where for each $v \in \{\text{l,r}\}$, $I_{j,p,q}^{(v,o)}(m,n) = I_j^{(v,o)}(m+p,n+q)$ and $(p,q) \in (\mathbb{Z}^*)^2$ refers to the spatial support of the neighboring pixels. Note that the size of this multivariate vector depends clearly on the number of the neighboring pixels used in each subband to take into account the spatial redundancies. Indeed, if we denote this number by N, the number of the components of the multivariate vector is given by $c = 2(N+1)$.

Then, the MGG distribution is used for modeling and the resulting parameters $\left(\alpha_j^{(M,o)}, \beta_j^{(M,o)}\right)$ are estimated using the moment matching method. Finally, the texture feature vector is built by taking the scalar coefficients of the estimated statistical parameters:

$$\mathbf{f}^{(M)} = \left(\alpha_j^{(M,o)}, \beta_j^{(M,o)}\right)_{\substack{o\in\{HL,LH,HH,LL\} \\ j\in\{1,\ldots,J\}}} \tag{3.14}$$

Once the joint texture statistics are extracted, the depth feature extraction technique and finally the 3D quality evaluation are performed similarly to the previous approach [54].

3.4 Results and Discussion

Unfortunately, to the best of knowledge, there is no available dataset for the quality assessment of stereo-laparoscopic images. For this reason, the proposed 3D IQA methods have been applied to the LIVE 3D Phase I database [154]. This database is composed of 365 symmetric stereo images, distorted by one of the five distortions namely blur, JPEG compression, JPEG2000 compression, fast fading and white noise. Amongst these distortions, noise is also relevant for stereo-laparoscopic images. Besides that, some of the other common distortions in stereo-laparoscopic images are uneven illumination, blur due to defocus and blur due to motion as illustrated in Figure 3.3. For the experiments on LIVE 3D Phase I database, two versions of multivariate statistical modeling have been considered by using $N = 3$ and $N = 8$ neighboring pixels, respectively. The wavelet representations of the stereo images as well as their associated disparity maps are generated by using two resolution levels. Figure 3.6 shows examples of

(a) reference left view

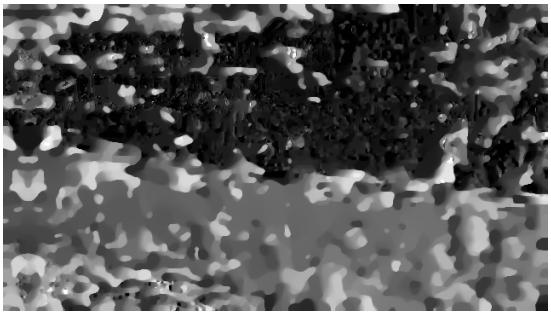
(b) referene right view

(c) FF distortion left

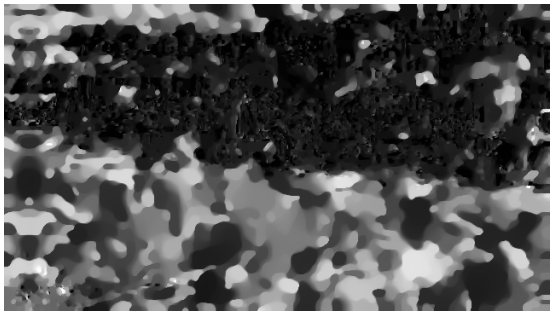
(d) FF distortion right

(e) JP2K distortion left

(f) JP2K distortion right



(g) Disparity FF stereopair



(h) Disparity JP2K stereopair

Figure 3.6: Reference, distorted laparoscopic stereo-pairs with along with their diparity maps from LIVE 3D Phase I

Table 3.1: LCC Comparison of Stereoscopic IQA for LIVE 3D Phase I Database

Method	JP2K	JPEG	WN	Blur	FF	Overall
Gorley [140]	0.485	0.312	0.796	0.852	0.364	0.451
Shen [139]	0.503	0.389	0.898	0.684	0.483	0.574
Benoit [141]	0.939	0.640	0.925	0.948	0.747	0.902
You [150]	0.877	0.487	0.941	0.919	0.730	0.881
Zhu [142]	0.807	0.379	0.517	0.777	0.503	0.626
Yang [151]	0.920	0.640	0.930	0.930	0.740	0.870
Hewage [152]	0.904	0.530	0.895	0.798	0.669	0.830
Akhter [153]	0.905	0.729	0.904	0.617	0.660	0.427
Chen [53]	0.907	0.695	0.917	0.917	0.735	0.895
Appina [52]	0.938	0.806	0.919	0.881	0.758	0.917
Hachicha [54]	**0.973**	**0.856**	**0.961**	**0.973**	**0.839**	**0.939**
BGG	**0.960**	0.815	0.956	**0.972**	**0.868**	**0.940**
MGG ($N = 3$)	0.954	0.794	**0.967**	0.964	0.804	0.928
MGG ($N = 8$)	0.957	**0.854**	**0.961**	0.970	0.807	0.932

distorted stereo-pairs from LIVE 3D Phase I database along with the reference image and their corresponding disparity maps. For the SVM and SVR tools used in the quality evaluation stage, we have employed 5-fold cross-validation with the radial basis function (RBF). The optimal values for C and γ have been found by making a search through sets $\{2^{-4}, 2^{-2}, \ldots, 2^{8}, 2^{14}\}$ and $\{2^{-6}, 2^{-4}, \ldots, 2^{4}, 2^{6}\}$, respectively of possible values.

The proposed approaches have been compared with various state-of-the-art methods presented in Section 5.1. The performance is evaluated using Pearson Linear Correlation Coefficient (LCC), Spearman's Rank-Ordered Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE). The obtained results are shown in Tables 3.1, 3.2 and 3.3, where the metrics of two best methods are highlighted in bold.

From the results, it can be firstly observed that the proposed joint statistical modeling-based 3D IQA approaches is among the two best methods. For instance, the bivariate statistics based method leads to the best results for fast fading distortion type with all of the metrics (i.e. LCC, SROCC and RMSE). Moreover, both the multivariate methods yield better results for white noise

Table 3.2: SROCC Comparison of Stereoscopic IQA for LIVE 3D Phase I Database

Method	JP2K	JPEG	WN	Blur	FF	Overall
Gorley [140]	0.420	0.015	0.740	0.749	0.366	0.142
Shen [139]	0.213	0.244	0.891	0.658	0.266	0.068
Benoit [141]	0.910	0.603	0.929	0.931	0.698	0.899
You [150]	0.859	0.438	**0.939**	0.882	0.588	0.879
Zhu [142]	0.770	0.292	0.465	0.793	0.475	0.639
Yang [151]	0.902	0.601	0.937	0.928	0.695	0.900
Hewage [152]	0.855	0.500	0.896	0.690	0.544	0.814
Akhter [153]	0.865	0.675	0.913	0.554	0.639	0.383
Chen [53]	0.863	0.617	0.919	0.878	0.652	0.891
Appina [52]	0.917	0.782	0.910	0.865	0.666	0.911
Hachicha [54]	**0.938**	**0.838**	0.932	**0.933**	**0.782**	**0.933**
BGG	0.912	0.785	0.926	**0.933**	**0.791**	**0.926**
MGG ($N = 3$)	**0.925**	0.774	**0.941**	**0.933**	0.726	0.918
MGG ($N = 8$)	0.921	**0.844**	0.932	**0.933**	0.706	0.925

and JPEG distortions. Furthermore, it can be noticed that the rest of results are comparable to the best of the scores obtained with all 3D IQA algorithms.

Finally, when we compare the results of our proposed methods with each other, we observe some interesting outcomes. The bivariate case gives better scores for the overall image set as well as for three distortions namely JPEG2000, blur and fast fading. On the other hand, the multivariate statistics with $N = 3$ neighboring coefficients gives the best results for white noise while the one with $N = 8$ neighboring pixels gives the best results for JPEG distortion. One reason that may explain this behavior of MGG method, for some distortions, could be because of disparity estimation. Indeed, the disparity estimation process depends on the quality of the left and right views. So, for some specific distortions, the disparity estimation method may fail and produce disparity maps with poor quality. This will certainly impact the disparity compensation process which is a crucial step in the MGG model (as shown in the Eq. 3.13).

Table 3.3: RMSE Comparison of Stereoscopic IQA for LIVE 3D Phase I Database

Method	JP2K	JPEG	WN	Blur	FF	Overall
Gorley [140]	11.323	6.212	10.197	7.562	11.569	14.635
Shen [139]	12.275	6.022	7.294	10.554	10.882	13.547
Benoit [141]	4.426	5.022	6.307	4.571	8.257	7.063
You [150]	6.206	5.709	5.621	5.679	8.492	7.747
Zhu [142]	7.681	6.068	14.720	9.127	10.736	12.782
Yang [151]	4.421	5.019	6.298	4.570	8.252	7.060
Hewage [152]	5.530	5.543	7.405	8.748	9.226	9.139
Akhter [153]	5.484	4.474	7.093	11.387	9.332	14.827
Chen [53]	5.402	4.523	6.433	5.898	8.322	7.247
Appina [52]	4.943	4.391	6.664	6.938	9.317	6.598
Hachicha [54]	**2.848**	**3.235**	4.481	**3.111**	**6.335**	**5.571**
BGG	**3.495**	3.585	4.689	**3.242**	**5.889**	**5.588**
MGG ($N = 3$)	3.738	3.757	**4.076**	3.575	7.058	6.076
MGG ($N = 8$)	3.656	**3.256**	**4.463**	3.313	6.949	5.893

3.5 Conclusion and Perspectives

This work focused on the use of joint statistics for no-reference quality assessment of stereoscopic images. These statistics are obtained using bivariate and multivariate generalized Gaussian distribution. The joint statistics aim to exploit the intra and inter-image dependencies. With these joint statistical texture features, we also added depth features extracted from the estimated disparity maps. Finally, the resulting features are used with a machine learning approach to predict the quality score of the stereo pairs.

The proposed methods tested with LIVE 3D Phase I database yield promising results and show the benefits of joint statistical modeling within a stereoscopic image for no-reference quality assessment. In the future, more dependencies can be incorporated to further improve the 3D quality prediction results. Moreover, other more efficient joint statistical models can also be investigated for a better discrimination between different distortion types. Lastly, it is very important to fulfill the need of quality assessment databases for medical images. This will be the objective of the next chapter while focusing on the laparoscopic videos data.

CHAPTER 4

New Subjective 2D Laparoscopic Video Quality (LVQ) Database

Abstract

There are a lot of challenges in designing an efficient metric for objective quality assessment of medical images and videos. One of the most challenging issue is to develop a metric consistent with the subjective evaluations. To accomplish this, one needs to have a benchmark database of images/videos, which has already been evaluated by human observers. For natural images and videos, many databases have been built. However, to the best of our knowledge, there is no such database of medical images/videos and especially laparoscopic data available publicly. For this reason, we have developed a new database of 2D laparoscopic videos as a first step. To this end, we have, in close cooperation with medical experts, identified the most common distortions affecting a laparoscopic video. Thereafter, we have collected multiple undistorted videos and have simulated these distortions at various severity levels to cover different possible scenarios. Once all the videos have been generated and finalized, their subjective quality has been evaluated by both medical expert and non-expert observers. The analysis of the obtained results provides an insight into the differences in observations between experts and non-experts [1][155].

[1] [155] **Khan, Z. A.**, Beghdadi, A., Alaya Cheikh, F., Kaaniche, M., Pelanis, E., Palomar, R., Fretland, Å. A., Edwin, B. and Elle, O. J., 2020, March. "Towards a video quality assessment based framework for enhancement of laparoscopic videos". In Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment (Vol. 11316, p. 113160P). International Society for Optics and Photonics.

4.1 Introduction and Motivations

A good video quality is an important requirement for ensuring optimal conditions for laparoscopic surgery. The distortions in a laparoscopic video not only affect the visibility of relevant structural information but also degrade the performance of subsequent computational tasks in robot-assisted surgery and image-guided navigation systems [156]. Examples of such tasks are segmentation [157, 158], instrument tracking [159, 160], and augmented reality [161].

Table 4.1: Common distortions affecting laparoscopic videos.

Distortions	**Likely Cause of Distortion**	**Possible Manual or Invasive Solutions**	**Computational Solutions**
Low or Bright Light [162]	Light source illumination problem	-Increase/Decrease lighting at light source [162] - Rectify light connection [162]	Enhancement Algorithms
Defocus Blurring [162]	Out of focus lens	- Correct focus [162] - Clean the lens [162]	Deblurring methods
Motion Blurring	Fast or continuous camera motion	- Move camera slowly - Keep camera stable	Deblurring methods
Noise (Graininess) [162]	Connection problems	Clean connection between camera head and laparoscope	Denoising methods
Uneven illumination	Light pointed sideways	Maintain light direction to the center	Enhancement algorithms
Smoke	Cauterization	Wait for smoke to clear	Desmoking Algorithms
Specular Reflection	Light reflections from tissue surfaces	Angle away from ROI	Specular reflection removal

Laparoscopic videos may be affected by different kinds of distortions during the surgery, resulting in visual discomfort and a loss of visual quality. These are mainly due to technical problems in the equipment [163] or side-effects of the surgical instruments (e.g. smoke with diathermy). To deal with such problems, most of the suggested prevalent solutions rely on making some changes to the technical equipment using one of the many available troubleshooting options [162]. However, all such solutions are time-consuming and hence require other alternatives especially in view of the patient safety. Moreover, existing solutions are also not flexible and do not always solve the problem at hand requiring eventually a specialist technician or a change of apparatus [163]. Table 4.1 summarizes the causes of commonly experienced distortions during laparoscopic surgery and their possible manual and computational solutions. From the table,

we can see that all manual or invasive solutions are either tedious, prone-to-errors and difficult to execute like for motion blur and uneven illumination or are disruptive and time-consuming like for most other distortions. Ideally, real-time computational solutions could help improve the overall surgical experience by countering the effects of these distortions.

In this chapter, we describe and propose a computational framework for automatic detection and correction of video quality for laparoscopic surgery (Figure 4.1). The proposed framework consists of a video quality assessment (VQA) module followed by an enhancement module. This work solely focuses on the video quality assessment part (dotted line in Figure 4.1) which is composed of two stages namely distortion detection/classification and video quality score evaluation. Such hybrid two-step quality assessment techniques are not new and have already been proposed for natural images [164].

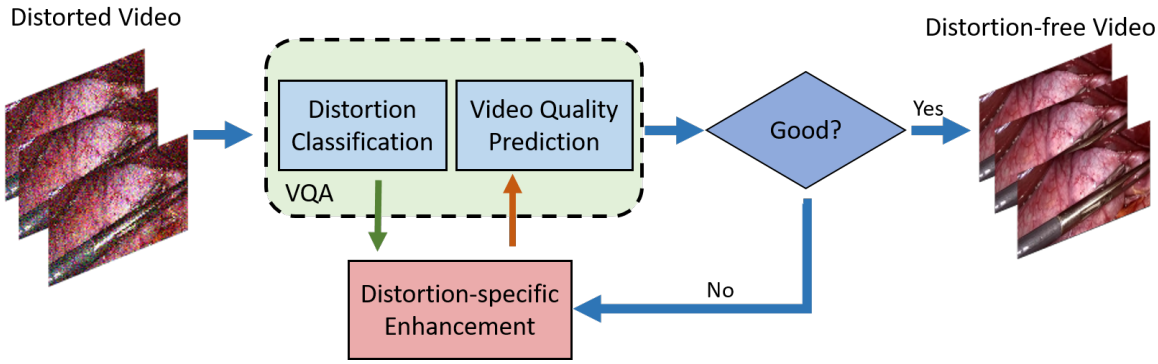


Figure 4.1: Basic Flow for Quality Monitoring Pipeline

Quality assessment of videos, if done subjectively, is time-consuming and hence not feasible in this context. In order to assess video quality automatically, objective metrics are needed. However, the effectiveness of an objective metric can only be evaluated by using a database of videos annotated with subjective scores [165]. Unfortunately, to the best of our knowledge, there is no such database of laparoscopic videos available publicly. Hence, there is currently a big gap to be filled in the field of medical visual quality assessment and especially in the surgical context. This work aims to fill this gap by proposing a new database which is dedicated to laparoscopic video quality assessment (Available at: LVQ Database) [1]

This chapter is organized as follows. Section 4.2 first provides a description of our new database. Then, Section 4.3 gives details on the selected reference videos while section 4.4 explains in detail the process of adding distortions to these selected videos. In Section 4.5, we have described the procedure of subjective testing. This is followed by a statistical analysis of

[1]URL: https://drive.google.com/file/d/1SoONeacp9vvihTY7zmWssG_cnVzx16oq/view?usp=sharing.

the obtained scores in Section 4.6. In Sections 4.7 and 4.8, we analyze the performance of some classifiers and objective quality metrics on our newly created database. Finally, discussions and concluding remarks are drawn in Section 4.9.

4.2 Description of the Video Quality Database

Our database called the Laparoscopic Video Quality (LVQ) database consists of a total of 10 reference videos, each of 10 seconds. The resolution of the videos is 512 × 288 with a 16:9 aspect ratio and a frame-rate of 25 fps. Each video is distorted by five different kinds of distortions with four different distortion strength levels, resulting in a total of 200 videos. Moreover, we have used uncompressed avi format for the videos to avoid introducing any kind of unwanted compression artefacts like blocking, blur or ringing. In the following sections, we describe in more details the construction of our database.

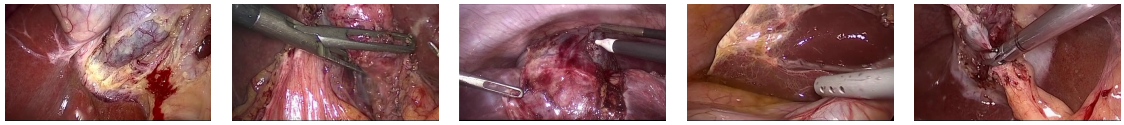
4.3 Selection of Reference Videos

For the database, ten different videos of laparoscopic cholecystectomy are selected as reference. These videos are extracted from Cholec80 dataset [166] and are shown in Figure 4.2. The selection of the videos is made with an attempt to include the maximum possible variations of the scene content and its temporal information. For scene content, ten different categories are chosen as illustrated in Figure 4.2. These are bleeding (BL), grasping and burning (GB), multiple instruments (MI), irrigation (IR), clipping (CL), stretching away (SA), cutting (CU), stretching forward (SF), organ extraction (OE) and burning (BU).

Furthermore, in order to analyze the diversity of videos we have used three parameters to characterize the videos [14] [165]. These are Spatial Information (SI), Colorfulness (CF) and Temporal Information (TI). Spatial Information provides a measure of edge energy. To find SI, each frame F_i is first filtered using the Sobel filter. The standard deviation of each filtered frame is then evaluated over the spatial domain. Finally, SI is given as the maximum of standard deviation values across all frames in a video [14] as depicted by Eq. (4.1)

$$SI = max_i\{std[Sobel(F_i)]\} \tag{4.1}$$

Colorfulness (CF) captures information related to variability and intensity of colors in a video. For a video represented in RGB format, CF can be computed using an opponent color space, defined using the components rg and yb where



(a) Bleeding (BL). **(b)** Grasp and Burn (GB). **(c)** Multiple instruments (MI). **(d)** Irrigation (IR). **(e)** Clipping (CL).

(f) Stretching away (SA). **(g)** Cutting (CU). **(h)** Grasping and stretching forward (SF). **(i)** Organ Extraction (OE). **(j)** Burning (BU).

Figure 4.2: One frame from each of the reference videos in the LVQ database.

$$\begin{aligned} rg &= R - G \\ yb &= 0.5(R+G) - B \end{aligned} \tag{4.2}$$

Then, CF is defined by the following equation:

$$CF = max_i\{\sqrt{\sigma_{rg}^2(F_i) + \sigma_{yb}^2(F_i)} + 0.3\sqrt{\mu_{rg}^2(F_i) + \mu_{yb}^2(F_i)}\} \tag{4.3}$$

where σ_x and μ_x for $x \in \{rg, yb\}$ represent standard deviation and mean of the pixel values in the color space for each frame.

Finally, Temporal Information (TI) measure is based on the motion difference feature, $M_i(m,n)$, which is the difference between pixel values located at same position (m,n) in two consecutive frames, F_i and F_{i-1}. It is given as:

$$M_i(m,n) = F_i(m,n) - F_{i-1}(m,n) \tag{4.4}$$

TI is then defined as the maximum of the spatial standard deviation of $M_i(m,n)$ over time:

$$TI = max_i\{std(M_i(m,n))\} \tag{4.5}$$

We evaluated these three parameters for all the selected reference videos in our database. Figure 4.3 shows the scatter plot of TI against SI whereas Figure 4.4 shows a plot of CF against SI. From the two plots, we can clearly observe that the selected videos for LVQ database contain diverse spatial information, colorfulness and temporal information.

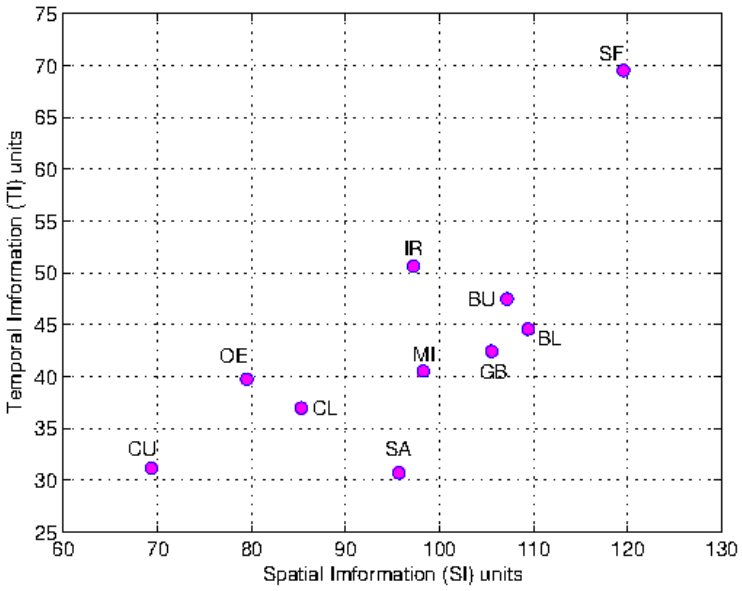


Figure 4.3: Plot of Temporal Information against Spatial Information for the selected videos.

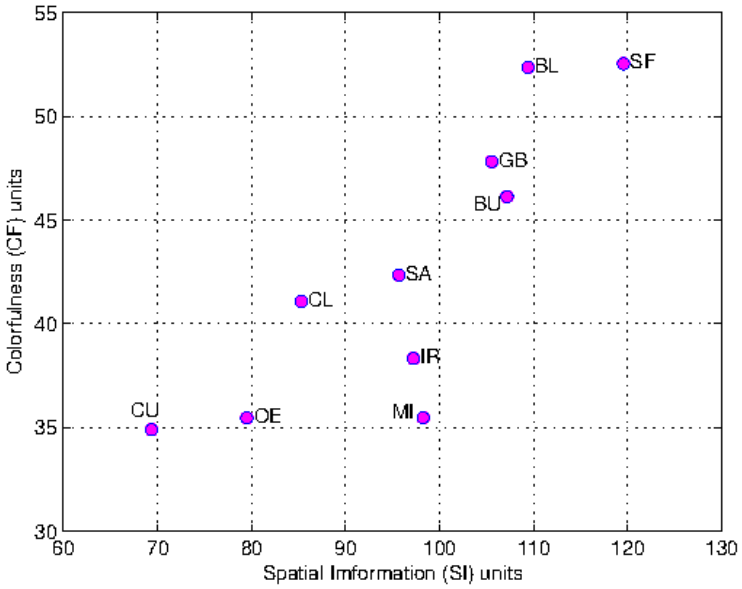
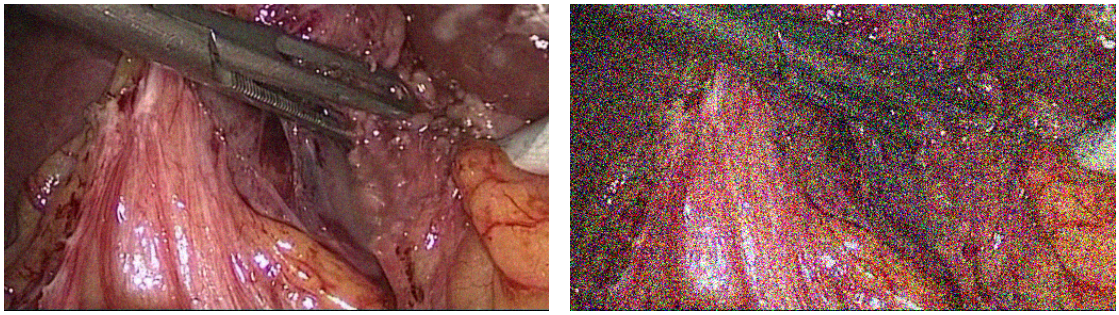


Figure 4.4: Plot of Colorfulness against Spatial Information for the selected videos.

4.4 Creation of Distorted Videos

We have chosen five different distortions for our database. These distortions, which are among the most common affecting a laparoscopic video, are the smoke, noise, uneven illumination, blur due to defocus and blur due to motion. In order to simulate each of these distortions, we have applied appropriate signal models for each distortion to each frame of the video. The details of



(a) (b)

Figure 4.5: One frame distorted by white Gaussian noise at (a) level 1 and (b) level 4.

modeling of each of these distortions are given below.

4.4.1 Additive White Gaussian Noise

Laparoscopic video may be affected by noise during acquisition due to camera sensors. Such noise can be modeled as additive white Gaussian noise as we have done for our database using the equation below for each frame $F_i(m,n)$.

$$d_i(m,n) = F_i(m,n) + N(m,n), \tag{4.6}$$

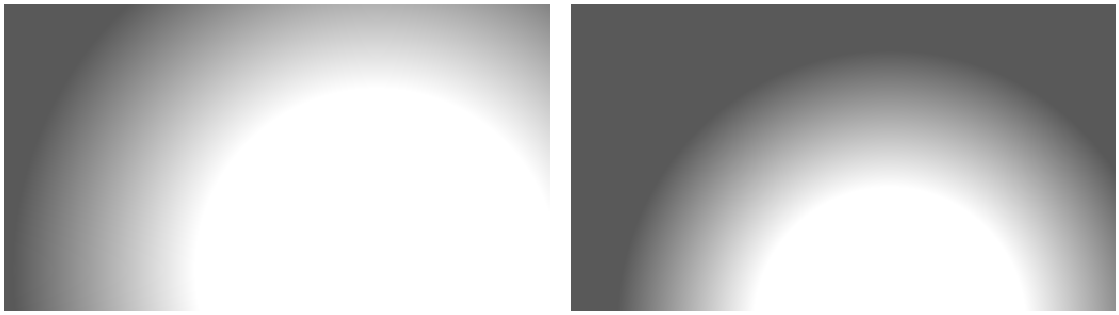
where $d_i(m,n)$ is the resulting distorted frame and $N(m,n)$ is the normally distributed random signal with the following probability density function:

$$p_N(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu_z)^2}{2\sigma^2}} \tag{4.7}$$

where μ_z is the mean of the distribution and σ stands its for standard deviation. In order to create four different levels, the variance of the Gaussian model was varied. The frames from two videos in the database with different levels of noise are shown in Figure 4.5.

4.4.2 Uneven Illumination

Laparoscopic video is often affected by uneven illumination where certain parts of the video are bright, being within the spot of the light from laparoscope, while the others are in the dark. In order to simulate this kind of distortion, we have initially generated a grayscale circular mask $C(m,n)$ having a bright circular region in the center and fading intensity towards the corners



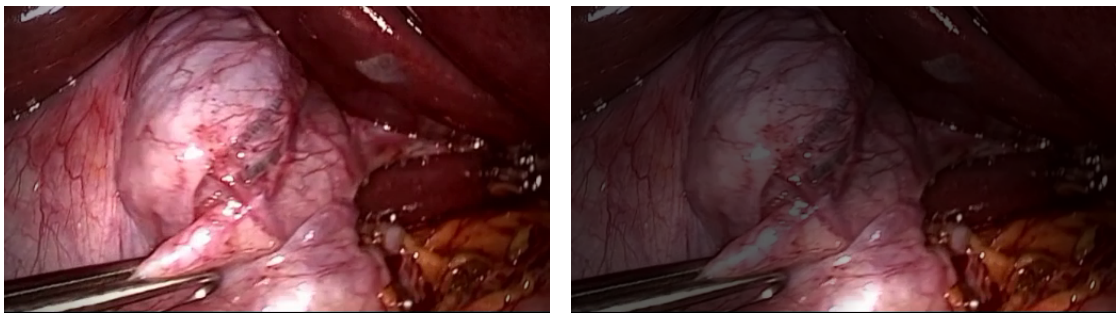
(a) **(b)**

Figure 4.6: Masks used to create two of the levels of uneven illumination in the LVQ database.

using the following equation

$$C(m,n) = \begin{cases} 1, & \text{for } r \leq H/AF \\ a, & \text{for } r \geq 2H/AF \\ 1 - \frac{(1-a)(r-H/AF)}{(H/AF)}, & \text{otherwise} \end{cases} \tag{4.8}$$

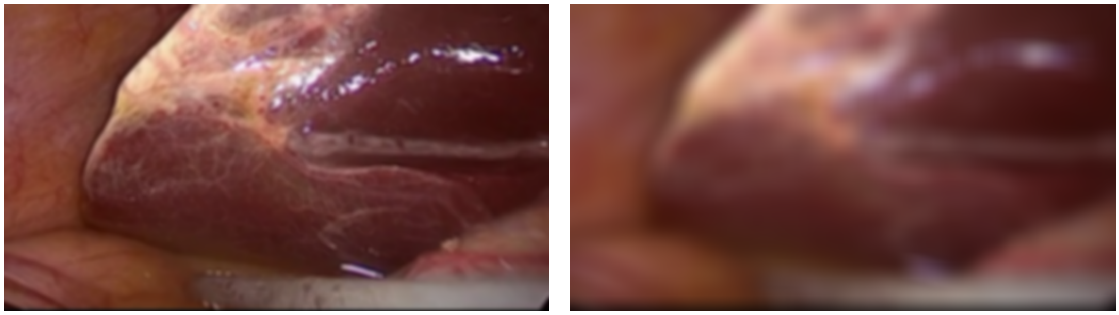
where H is the height of the image, AF is the area factor, a is the attenuation and r is the distance of point (m,n) from the center of the circular regions. In real scenarios with uneven illumination, the bright region is often not in the center of the frame but rather pointed to one of the sides. For this reason, we have chosen the center of the mask to be slightly sideways to create a realistic effect. Figure 4.6 shows two of the masks generated using Eq. (4.8) with each mask representing a different severity level.



(a) **(b)**

Figure 4.7: One frame distorted by uneven illumination at (a) level 1 and (b) level 4.

Using the generated mask, uneven illumination is then added to the original frame $F_i(m,n)$



(a) (b)

Figure 4.8: One frame distorted by defocus blur at (a) level 1 and (b) level 4.

using simple pixel-by-pixel multiplication to give the distorted frame $d_i(m,n)$

$$d_i(m,n) = C(m,n)F_i(m,n), \tag{4.9}$$

In order to create the four levels of this distortion, we modify both the center point of the mask and the area of the central region. Figure 4.7 illustrates a video frame with uneven illumination distortion applied.

4.4.3 Blur due to Defocus

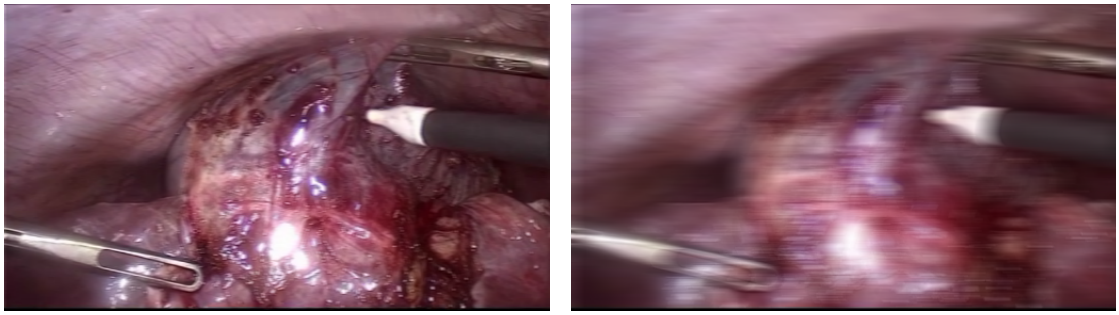
Blur caused by an out-of-focus lens is a very common problem and can affect the performance of a laparoscopic procedure. In order to simulate this, an isotropic low-pass Gaussian filter is applied to each frame using

$$d_i(m,n) = h(m,n) * F_i(m,n), \tag{4.10}$$

where $F_i(m,n)$ and $d_i(m,n)$ are the original and filtered frames respectively while $h(m,n)$ is the Gaussian filter defined by

$$h(m,n) = \frac{1}{2\pi\sigma^2} e^{-\frac{m^2+n^2}{2\sigma^2}} \tag{4.11}$$

where σ is the standard deviation of the Gaussian distribution. To generate different levels, we have used different values of size and the standard deviation for the filter. Figure 4.8 illustrates video frames with defocus blur applied.



(a) (b)

Figure 4.9: One frame distorted by motion blur at (a) level 1 and (b) level 4.

4.4.4 Blur due to Motion

Due to frequent and random motion occurring during the laparoscopic surgery, often times blurriness due to motion can be observed, causing visual discomfort to the surgeon. For our LVQ database, this distortion was generated using motion filter implementation of MATLAB. In essence, the motion blur is added by convolving the original frame F_i with the linear shift-invariant point spread function (PSF) based motion filter, h_M

$$d_i(m,n) = h_M(m,n) * F_i(m,n), \tag{4.12}$$

We have simulated horizontal blur only and varied the length of the filter to generate different levels of distortion. Figure 4.9 illustrates video frames with motion blur at two different severity levels.

4.4.5 Smoke

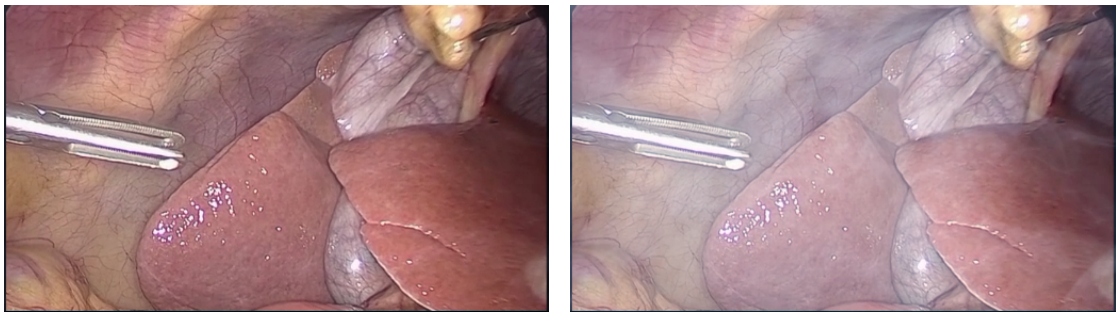
Smoke is a common distortion encountered during laparoscopy and occurs during the activities where tissue is burnt using cautery instrument. It is a difficult distortion to synthesize for videos. In order to generate videos with smoke, we have made use of a video containing only the smoke with a black background as illustrated in Figure 4.10. This video is then blended with the reference video using screen blending mode to create the effect of smoke in the video. We can therefore add the smoke in a reference frame using the following equation:

$$d_i(m,n) = 1 - (1 - F_i(m,n))(1 - \alpha S_i(m,n)), \tag{4.13}$$

where $F_i(m,n)$ and $d_i(m,n)$ are the original and the resulting distorted frames respectively, α is the opacity and $S_i(m,n)$ is the smoke video frame at time i.

Figure 4.10: Single frame from smoke-only video with black background.

By adjusting the value of opacity α, we generate different levels of smoke in the video. Figure 4.11 illustrates frames from videos affected by two different levels of smoke.



(a) **(b)**

Figure 4.11: One frame distorted by smoke at (a) level 1 and (b) level 4.

4.5 Subjective Testing

For the subjective testing, we have used pairwise-comparison protocol [14, 78]. For each observer, we randomly displayed all possible pair combinations of distorted videos, such that each pair had two videos from the same category and the same distortion type, with only difference being the severity level. This corresponded to 6 pair-wise comparisons per reference video for each distortion type as can be found using the following equation where $L = 4$ and corresponds to the

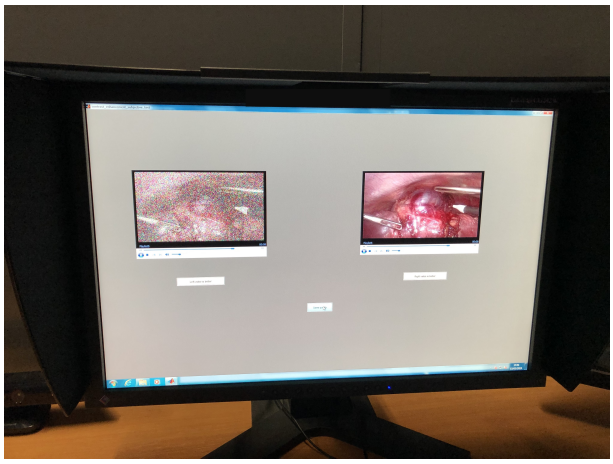


Figure 4.12: Setup for subjective tests

number of severity levels for video i affected by distortion type D:

$$N_{PC}^{D}(i) = {}^{L}\mathbf{C}_2 = \frac{L!}{2!(L-2)!} \tag{4.14}$$

The observers had the choice to give an equal score to the videos if they perceived. For each comparison, the preferred video was given one point. In case of an equal choice, a score of 0.5 was given to each displayed video. The observers were shown each video once and they had the possibility to see the video again if they wanted. At the end, the scores for each video from all the observers were added. Finally, the Mean Opinion Score (MOS) for the i-th video was obtained by averaging the total score for that video over number of observers N_o [167].

$$MOS_i = \frac{1}{N_o}\sum_{j=1}^{N_o} score_{ij} \tag{4.15}$$

In order to perform the subjective tests, a calibrated 24.1 inch LCD monitor was used. The observers were forced to perform the experiments at a fixed distance of twice the screen height which is equivalent in our case to be 4.5 times the image height of the image. The setup for the subjective tests is shown in Figure 4.12. All the observers had either normal vision or corrected to normal vision and they went through a pre-screening procedure for color vision and visual acuity.

In total, thirty non-expert and ten expert observers performed the subjective tests for the database. Both the expert and non-expert observers were considered as two separate groups. For each group, outliers were first detected based on non-transitivity. This corresponded to one subject in each group. The preference matrices for remaining subjects in the two groups were then

Table 4.2: Sample Preference Matrices for a video with defocus blur aggregated over preferences of 29 non-expert observers (left) and 9 expert observers (right)

–	L_1	L_2	L_3	L_4	p_i
L_1	0	27.5	29	29	85.5
L_2	1.5	0	22.5	29	53
L_3	0	6.5	0	28.5	35
L_4	0	0	0.5	0	0.5

–	L_1	L_2	L_3	L_4	p_i
L_1	0	6	8	9	23
L_2	3	0	5.5	9	17.5
L_3	1	2.5	0	8.5	12
L_4	0	0	0.5	0	0.5

compiled by aggregating preferences of the observers. In Table 4.2, we can see sample preference matrices obtained from experts and non-experts for the same video affected by different levels of defocus blur. It can be clearly seen from the table that as the severity level increases, the aggregate score p_i decreases.

4.6 Statistical Analysis of Scores

In this section, we analyze the recorded scores. First, in order to evaluate the reliability and validity of the collected subjective data, we evaluate the coefficient of agreement for the two groups. Thereafter, we analyze the intra-rater agreement to check for inconsistencies. Finally, we assess the differences between the scores of experts and non-experts.

4.6.1 Inter-rater Reliability

We evaluate the inter-rater reliability using Kendall's coefficient of agreement, u [168]. For the pair comparison, it is given by:

$$u = \frac{2\sum_{j=1}^{K}\binom{p_{ij}}{2}}{\binom{N_o}{2}\binom{L}{2}} - 1 \tag{4.16}$$

where L is the number of severity levels, N_o is the number of observers and p_{ij} represents the number of times video with severity level i is preferred to video with level j. The value of Kendall's coefficient of agreement is 1 when all observers agree on their preferences.

For testing significance of coefficient of agreement, we have performed the chi-squared test (χ^2). The χ^2 values are evaluated using Eq. (4.17)

$$\chi^2 = \frac{L(L-1)(1+u(N_o-1))}{2} \tag{4.17}$$

The degree of freedom χ^2 is selected as $\frac{L(L-1)}{2}$. The minimum value of u is $\frac{-1}{(N_o-1)}$ and $\frac{-1}{N_o}$ for even and odd number of observers respectively. For our experiment, for 30 non-expert

Table 4.3: Coefficient of Agreement for **non-experts** for all distortions in all videos

video	$u_{defocus}$	$\chi^2_{defocus}$	u_{noise}	χ^2_{noise}	u_{illum}	χ^2_{illum}	u_{smoke}	χ^2_{smoke}	u_{motion}	χ^2_{motion}
1	0.978	170.28	0.823	144.21	0.715	126.14	0.854	149.45	0.896	156.48
2	0.978	170.28	0.872	152.48	0.902	157.45	0.796	139.65	0.879	153.72
3	0.954	166.28	0.875	152.03	0.777	136.48	0.750	132.07	0.849	148.62
4	0.978	170.28	0.856	149.86	0.839	146.97	0.850	148.76	0.934	162.97
5	0.978	170.28	0.894	156.21	0.914	159.52	0.744	130.97	0.913	159.38
6	0.912	159.24	0.906	158.14	0.688	121.59	0.700	123.66	0.785	137.86
7	0.977	170.14	0.977	170.14	0.890	155.52	0.852	149.17	0.895	156.34
8	0.869	152.07	0.913	159.38	0.869	151.93	0.737	129.86	0.814	142.83
9	0.956	166.55	0.978	170.28	0.955	166.41	0.891	155.66	0.898	156.90
10	0.956	166.55	1.000	174.00	0.892	155.93	0.818	143.38	0.921	160.76

observers, the minimum value of the consistency coefficient u_{min} is $\frac{-1}{29} = -0.0345$ and for 10 expert observers it is -0.111. The null hypothesis H_0 is rejected when the observed χ^2 is greater than its critical value. Tables 4.3 and 4.4 show the values of u and χ^2 for non-experts and experts respectively.

4.6.2 Intra-rater Agreement

In order to assess intra-rater agreement, we have further evaluated coefficient of transitivity or consistency, τ. It is measured based on the number of circular triads or intransitivity in a ranked data. The coefficient of consistency can be calculated using the relation [168]:

$$\tau = \left\{ \begin{array}{ll} 1 - \frac{24d}{L^3 - L}, & \text{if } L \text{ is odd,} \\ 1 - \frac{24d}{L^3 - 4L}, & \text{if } L \text{ is even} \end{array} \right\} \tag{4.18}$$

where d is the number of circular triads. For each video, we evaluated the coefficient of consistency by averaging the coefficients across all the observers and found no inconsistencies.

4.6.3 Comparison of Expert and Non-Expert Scores

In order to gain an insight into how differently medical experts observe surgical videos, we perform an analysis of the differences in subjective scores from experts and non-experts. First of all, we take a look into the way experts and non-experts perceive different severity levels of each distortion. For this, we first evaluate a mean normalized score, s_{norm} for all the 10 videos

Table 4.4: Coefficient of Agreement for **experts** for all distortions in all videos

video	$u_{defocus}$	$\chi^2_{defocus}$	u_{noise}	χ^2_{noise}	u_{illum}	χ^2_{illum}	u_{smoke}	χ^2_{smoke}	u_{motion}	χ^2_{motion}
1	0.926	50.44	0.738	37.00	0.774	38.50	0.726	36.50	0.929	45.00
2	1.000	54.00	0.857	42.00	0.738	37.00	0.679	34.50	0.857	42.00
3	0.852	46.89	0.929	45.00	0.774	38.50	0.667	34.00	0.821	40.50
4	0.852	46.89	0.857	42.00	0.691	35.00	0.917	44.50	1.000	48.00
5	0.861	47.33	0.929	45.00	0.643	33.00	0.774	38.50	0.786	39.00
6	0.741	41.56	0.857	42.00	0.655	33.50	0.679	34.50	0.798	39.50
7	0.806	44.67	0.857	42.00	0.762	38.00	0.786	39.00	0.786	39.00
8	0.648	37.11	0.929	45.00	0.631	32.50	0.548	29.00	0.560	29.50
9	0.926	50.44	0.917	44.50	0.738	37.00	1.000	48.00	0.917	44.50
10	0.833	46.00	0.917	44.50	0.786	39.00	0.786	39.00	0.810	40.00

affected by the same type of distortion D_j at the same severity level L_i.

$$s_{norm}(D_j, L_i) = \frac{1}{10} \sum_{n=1}^{10} MOS_n(D_j, L_i) \tag{4.19}$$

Figure 4.13 shows a comparison between expert and non-expert mean normalized scores for LVQ database. From the figure, we can clearly see how experts perceive quality differently for all distortions except for defocus blur. The difference is more pronounced for less distorted videos (levels 1 and 2) suggesting how even the slightest level of distortion affects the perception of a video for experts (who are more task-oriented).

To further analyze the differences, we look into the variations of MOS with respect to the distortion severity level. For this, we have plotted boxplots in Figure 4.14 for data from both the experts and the non-experts. These plots show how there are large variations in the opinions of experts as compared to non-experts. This is one indicator highlighting the difficulty in modeling the task-oriented visual perception in the medical context.

Finally, we further plot a scatter plot shown in Figure 4.15 to compare the spread of MOS for experts and non-experts. From the plot, we can see a higher spread of MOS in the vertical direction (experts) as compared to the horizontal direction (non-experts) highlighting the difference in opinion between the experts.

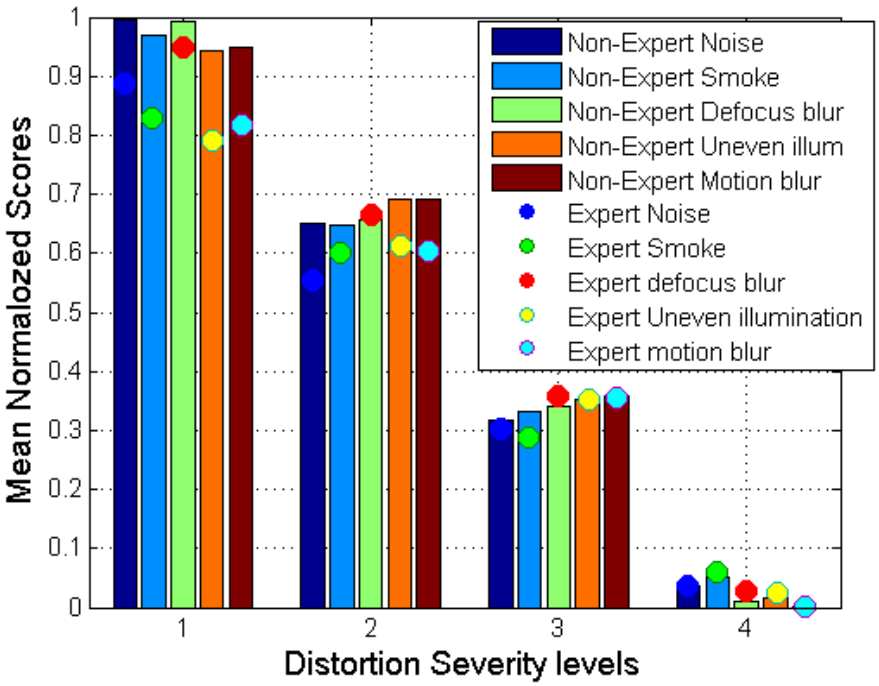
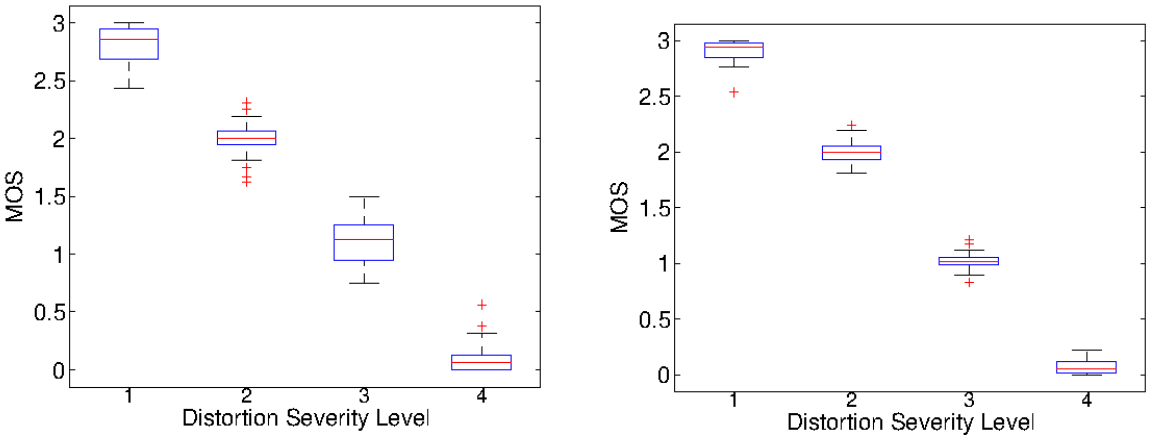


Figure 4.13: Comparison of subjective scores for experts and non-experts.



(a) **(b)**

Figure 4.14: Boxplots for (a) expert and (b) non-expert scores.

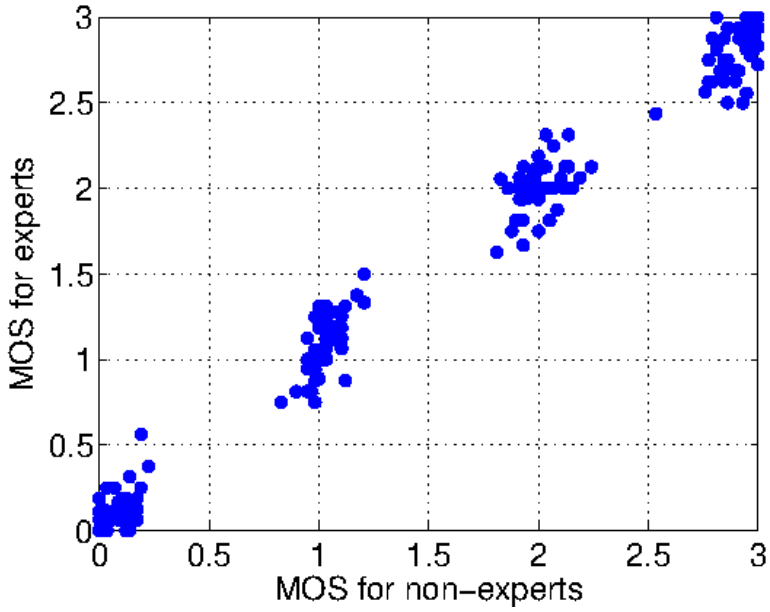


Figure 4.15: Expert vs non-expert scatter plot.

4.7 Distortion-specific Classifiers

In order to apply the appropriate enhancement method with suitable parameters in our proposed video enhancement framework, VQA module should be able to detect the type of distortion affecting a video as well as its severity level. For this reason, our VQA consists of a distortion identification step followed by the quality score estimation.

As a first step, for each kind of distortion, we have chosen a distortion-specific classification method. In our selection of these methods, we have gone for no-reference, opinion-unaware, accurate and computationally less expensive methods to allow for real-time performance. Below are the details of the classification methods used for each kind of distortion.

4.7.1 Motion and defocus blur

For the two kinds of blur, we have used Perceptual Blur Index (PBI) [87] with threshold as the classifier. PBI is a quality metric for estimating the level of blur in an image. It is based on the way Human Visual System (HVS) perceives addition of blur to an already blurred image and to a sharp one differently. The perceptual difference is more pronounced for the latter case. It is defined in terms of the difference between total radial energy of the input image $RE(w)$ and that of its binomial filtered version $RE_f(w)$ as

$$PBI = \log(\frac{1}{w_{max}} \sum_{w} |RE(w) - RE_f(w)|) \tag{4.20}$$

where w_{max} is the maximal frequency.

4.7.2 Smoke

In order to detect if there is smoke in a video, we have used Saturation Analysis (SAN) classifier [169]. SAN classifier uses the histogram of saturation channel of a frame to detect smoke. If the majority of bin values in histogram $hist$ are below the chosen threshold t_c, the video frame is classified to have smoke in it. The threshold used is $t_c = 0.35$ as suggested in the original work [169]. The probability of an image having smoke $p_{(S)}$ and no smoke $p_{(NS)}$ are therefore defined as

$$p_{(S)} = \frac{1}{|hist|} \sum_{\substack{i=0 \\ b \in hist \\ t \leq t_c}} b_i \tag{4.21}$$

$$p_{(NS)} = 1 - p_{(S)} \tag{4.22}$$

where b_i is the i-th bin value of the histogram $hist$.

4.7.3 Noise

For noise classification, we have chosen the fast noise variance estimator [170] with threshold. In this method, the standard deviation of additive white Gaussian noise in an image is estimated using a noise estimation mask. This suggested mask, M_N has been generated using a difference of two 3×3 masks, each approximating the Laplacian of an image. For an image I with width W and height H, the estimated standard deviation σ_n of noise is given by [170]:

$$\sigma_n = \sqrt{\frac{\pi}{2}} \frac{1}{6(W-2)(H-2)} \sum_{m,n} |I(m,n) * M_N| \tag{4.23}$$

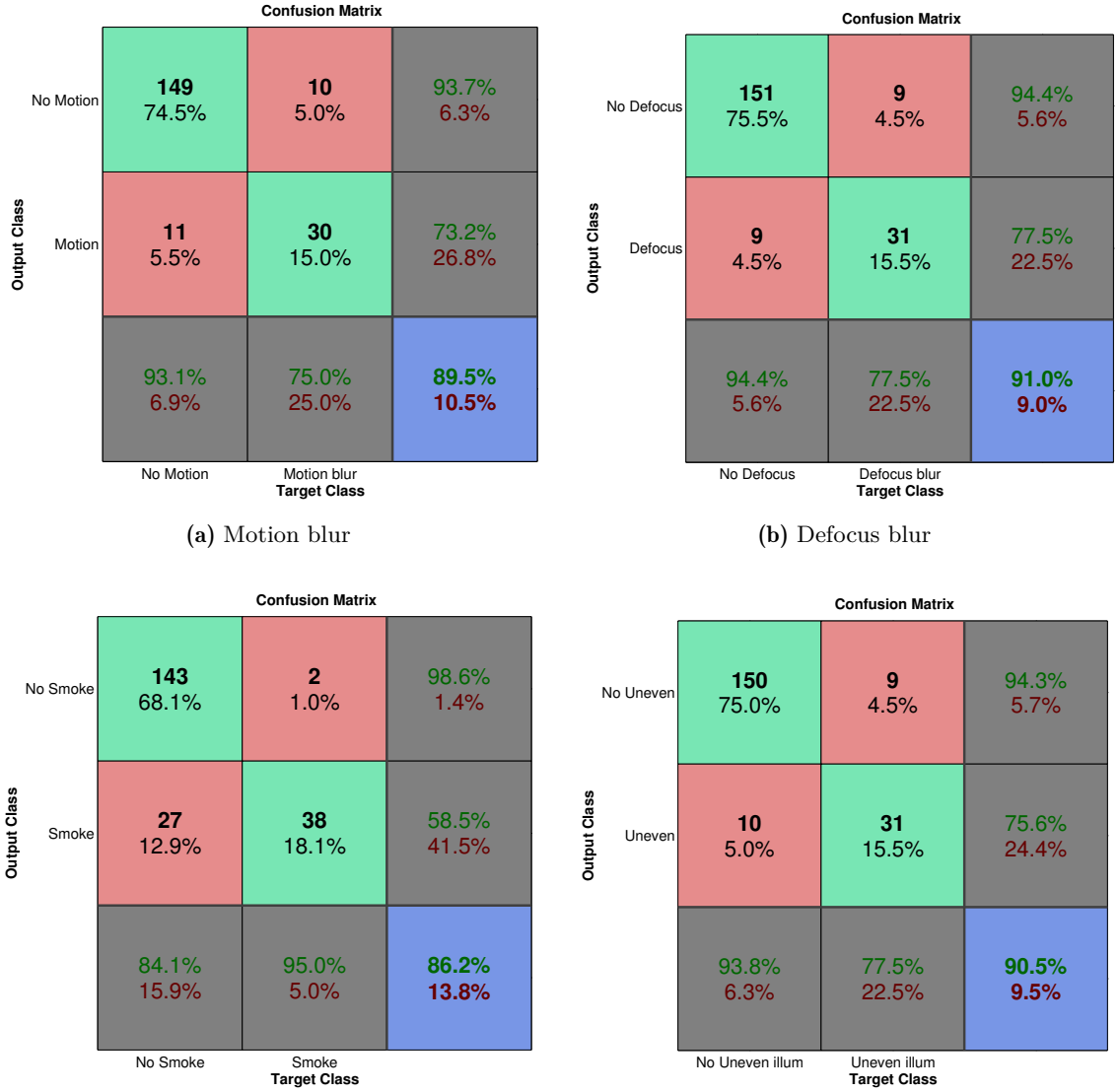
4.7.4 Uneven illumination

In order to detect whether a video is affected by uneven illumination or not, we have developed a novel classifier which makes use of statistics of the luminance component of an image. For an unevenly illuminated laparoscopic video frame, there are some dark regions in the image which tend to increase the range of values for the luminance component in an image, while reducing the mean luminance value of the image at the same time. Making use of these trends, we have proposed a new classifier that uses a threshold on the Luminance Mean to Range (LMR) ratio, defined simply as the ratio of the mean luminance value to that of the range of luminance values in an image. For an image with N_p pixels and with luminance component Y, this index can be

defined as

$$LMR = \frac{\frac{1}{N_p}\sum_{i=0}^{N_p} Y_i}{max(Y) - min(Y)} \tag{4.24}$$

An image with a LMR value smaller than a pre-defined threshold can be classified to have been affected by uneven illumination.



(a) Motion blur

(b) Defocus blur

(c) Smoke

(d) Uneven illumination

Figure 4.16: Confusion matrices from classifiers for LVQ database

To evaluate the performance of our selected classification methods, all the videos in our database were passed through these classifiers. The results obtained for classification accuracies were: 86.2% for smoke; 89.5% for motion blur; 91% for defocus blur; 100 % for noise and 90.5% for uneven illumination classifiers. Figure 4.16 shows the confusion matrices for 4 distortions other than noise. From these matrices, we can observe that only the smoke classifier gives a

relatively poor performance with a high number of false negatives (around 13%), whereas all other classifiers perform well with both precision and recall values of higher than 0.93.

4.8 Video Quality Score

Once the distortion is identified in a video, we can evaluate its severity using a quality score. To this effect, we have selected 3 different metrics which are often used as benchmarks for natural images and videos. These are PSNR, SSIM[23] and VIF[171]. However, for laparoscopic videos, there is usually no ground truth available and so a no-reference (NR) metric makes more sense. For this reason, we have also included NR metrics. However, due to the limited number of good NR metrics for videos, we have only selected one of the more recent NR metrics dedicated to opinion-unaware VQA called VIIDEO[36]. We have also included two NR image quality metrics BRISQUE[35] and NIQE[172]. For both of these, we have used the mean metric value from all frames as the score for the video.

Table 4.5: PLCC for **non-expert** scores in LVQ Database (best two values in bold for each column)

Metric	Noise	Defocus Blur	Motion Blur	Uneven illumination	Smoke	Overall
PSNR	**0.9968**	0.8166	0.8199	0.9561	**0.9811**	0.6054
SSIM	0.9690	0.7388	**0.8861**	**0.9926**	0.9165	**0.6123**
VIF	**0.9925**	**0.9764**	**0.9713**	**0.9919**	**0.9853**	**0.6267**
BRISQUE	0.9803	0.9646	0.4090	0.3142	0.3735	0.4593
NIQE	0.9783	**0.9880**	0.7704	0.6618	0.3238	0.4242
VIIDEO	0.8749	0.3549	0.4998	0.3983	0.4214	0.3842

In order to assess whether an existing objective video quality metric correlates well or not with subjective scores, Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) were evaluated for the metric scores after performing a non-linear regression with a 5-parameter logistic function. From Tables 4.5 and 4.7, we can see that none of the objective metrics perform well when overall correlations are evaluated, with maximum PLCC value of 0.6267 with VIF for non-experts and a value of 0.6853 with PSNR for experts.

However, with individual distortion types, VIF correlates much better with subjective scores

Table 4.6: SROCC for **non-expert** scores in LVQ Database (best two values in bold for each column)

Metric	Noise	Defocus Blur	Motion Blur	Uneven illumination	Smoke	Overall
PSNR	0.9594	0.7773	0.8163	0.9372	**0.9439**	0.5775
SSIM	0.9509	0.7157	**0.8941**	**0.9502**	0.8987	**0.5914**
VIF	**0.9636**	**0.9417**	**0.9433**	**0.9391**	**0.9316**	**0.6228**
BRISQUE	0.9571	0.9332	0.3564	0.2980	0.4041	0.4304
NIQE	**0.9640**	**0.9514**	0.6101	0.5416	0.3589	0.3731
VIIDEO	0.8600	0.3138	0.379	0.3888	0.3866	0.3416

Table 4.7: PLCC for **expert** scores in LVQ Database (best two values in bold for each column)

Metric	Noise	Defocus Blur	Motion Blur	Uneven illumination	Smoke	Overall
PSNR	**0.9939**	0.8146	0.8226	0.9452	**0.9777**	**0.6853**
SSIM	0.9706	0.7358	**0.8827**	**0.9847**	0.9116	0.5732
VIF	**0.9896**	**0.9806**	**0.9708**	**0.9878**	**0.9808**	**0.5909**
BRISQUE	0.9761	0.9623	0.4208	0.2973	0.4009	0.4434
NIQE	0.9741	**0.9883**	0.7836	0.6655	0.4301	0.4407
VIIDEO	0.8658	0.3498	0.5136	0.4035	0.4195	0.3744

as compared to others for both groups and for all the distortions. Among the NR metrics, both NIQE and BRISQUE give good results for noise and defocus blur, with NIQE being the better of the two for motion blur and uneven illumination. However, VQA specific method VIIDEO performs poorly for all distortions except for the noise.

All these results are very significant as they imply that none of these metrics are generic or non-distortion specific for the kind of videos and distortions encountered in the medical domain. Moreover, these results also show a difference with respect to their correlation with expert and non-expert scores. To be more specific, if we compare the results of experts and non-experts, we can see that generally all the metrics tend to correlate better with non-expert opinion as compared to expert opinion.

Table 4.8: SROCC for **expert** scores in LVQ Database (best two values in bold for each column)

Metric	Noise	Defocus Blur	Motion Blur	Uneven illumination	Smoke	Overall
PSNR	0.9579	0.7836	0.7977	0.9530	**0.9478**	**0.6914**
SSIM	0.9435	0.7320	**0.8802**	**0.9580**	0.8817	**0.5653**
VIF	**0.9592**	**0.9555**	**0.9376**	**0.9534**	**0.9459**	0.5642
BRISQUE	0.9527	0.9355	0.3994	0.2634	0.4355	0.3842
NIQE	**0.9594**	**0.9443**	0.7028	0.5605	0.3382	0.3674
VIIDEO	0.8822	0.3023	0.3915	0.4281	0.4416	0.3334

4.9 Discussions and concluding remarks

In this chapter, we have proposed a novel computational framework for laparoscopic video enhancement based on video quality assessment. Especially, we have taken a major initiative for quality assessment of laparoscopic videos by creating a database with subjective quality scores not only from normal observers but also from medical experts. Our initial results show that the existing NR metrics for video quality assessment are not sufficient especially in context of laparoscopic videos. Moreover, we have observed that experts and non-experts differ in their opinions on video quality assessment and new no-reference metrics are required to model expert opinion. In this regards, the constructed LVQ database is an important step to address the above challenges in a future work and in facilitating development of new VQA metrics in the medical imaging context. We address this issue of lack of a good objective quality assessment metric for medical data in the next chapter.

CHAPTER 5

Residual Network based No-Reference Video Quality Assessment for 2D Laparoscopic Videos

Abstract

Objective Video Quality Assessment (VQA) is a very challenging problem. For applications where there is a need to continuously monitor and enhance the video quality, VQA method must not only predict the level of deterioration but also detect the type of distortion affecting the video. Conventional VQA methods generally do not perform well for all kinds of videos and distortions. On the other hand, deep learning methods have better generalization performance while requiring large amounts of labeled data. Furthermore, they also do not tackle the problem of distortion classification and only focus on quality prediction part. In this work, we propose a new Residual network based VQA method which not only predicts the quality score but also detects the type of distortion affecting the video. Furthermore, to tackle the problem of limited data, we have proposed to use a ranking based pre-training approach followed by transfer learning. Results on a laparoscopic video quality (LVQ) dataset show that our proposed approach outperforms other state-of-the-art VQA methods [1][155].

[1] [173] **Khan, Z.A.**, Beghdadi, A., Kaaniche, M. and Alaya Cheikh, F., 2020, October. "Residual Networks Based Distortion Classification and Ranking for Laparoscopic Image Quality Assessment". In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 176-180). IEEE.

5.1 Introduction

Monitoring and assurance of a good video quality is a very critical task in all kinds of applications like video streaming, video surveillance, medical endoscopy and underwater exploration [174, 175]. Any loss of useful information in these videos, resulting from some kinds of video distortions, may not only affect the visual experience but could also cause fatal consequences like in the case of surveillance or medical videos. Hence, it is necessary to timely detect the distortion and correct the affected video using a given enhancement technique. The causes of these distortions are manifold but can broadly be divided into two types. The first type of distortion is introduced from within the video recording and transmission system [163]. The examples of such distortions are noise, defocus blur and compression artifacts. The second type of distortions result from the scene content being visualized. Examples of such distortions are smoke, non-uniform illumination and specular reflections. No matter what the cause of the distortion is, a good video showing the visual content clearly is the basic requirement of any video based system.

In order to assess the video quality, some metrics have already been proposed in literature [176]. These Video Quality Assessment (VQA) metrics are classified into Full-Reference, Reduced-Reference or No-Reference. Full-Reference metrics evaluate the quality of a video in relation to a pristine or a reference video. On the other hand, in Reduced-reference quality evaluation, only some information related to the reference video is available and used. However, in many real applications, no reference video or its related information is readily available. For this reason, No-Reference (NR) metrics are needed. The focus of this work is also on NR-VQA metric.

Since a video is a sequence of frames, VQA has always closely been related to its image counterpart or Image Quality Assessment (IQA). IQA has received more attention as compared to VQA over the years and initially NR-IQA was extended to NR-VQA by applying it to each frame before averaging the results from all frames. However, with the passage of time, video-specific metrics have been proposed in addition to more efficient temporal pooling strategies than the simple average pooling method [177].

Most conventional VQA metrics make use of hand-crafted features from videos. This makes these metrics less robust in the sense that they work well only for the types of distortions for which they are designed and not as well for other kinds of distortions. To overcome this drawback, some works have recently employed deep neural network architectures for VQA. However, one major challenge in training these deep networks is the requirement of large amounts of labeled video data. In the context of VQA, these labels are in the form of subjective quality scores obtained from evaluations of the videos by human observers [14]. Large amounts of such evaluations are

time-consuming and hence hard to undertake. For this reason, most existing public video quality databases have a limited number of videos [165]. To overcome this lack of data, either data augmentation technique is used or the video data is divided into frames or patches to increase the size of the training data. However, the scores derived for the patches or frames may not always be reliable and perceptually relevant.

In this work, we propose a Residual networks based method for VQA which is first trained on ranked data before fine-tuning it with the labeled data. This approach helps to overcome the issue of the lack of data since the ranked data could be generated easily and in large amounts without the need of subjective testing. A similar method has been proposed for IQA that has been named RankIQA [178]. However this work is significantly different in two aspects. First, unlike RankIQA, in the first step the network is trained for dual tasks of distortion classification and ranking instead of classifying the ranks only. This is important since the knowledge of the type of distortion is as significant as knowing about the level of its severity for subsequent enhancement. Secondly, the temporal aspect is added to the network so that instead of deriving the scores for frames from videos, this aspect is included within the network allowing for the design of an end-to-end framework for VQA.

From existing works, we can identify two major challenges in the design of an efficient VQA metric based on deep learning. First, there is a need to overcome the problem of limited labeled training data for videos. Secondly, there is a need for an end-to-end trainable network which considers both spatial and temporal aspects and works well for different kinds of distortions simultaneously. In this work, we have tried to solve these two problems for quality monitoring applications. The rest of this chapter is organized as follows. In section 5.2, we first describe the current state-of-the-art on deep learning for image and video quality assessment. Then, in Section 5.3, we present some details about Residual networks. Thereafter, in Section 5.4 we introduce our proposed architecture for distortion classification and ranking in detail. This is followed by the details about our proposed approach for laparoscopic video quality prediction in Section 5.5. Then in Section 5.6, we describe our experiments, results and discussion. Finally, in the last section, we present our conclusions and future perspectives.

5.2 Deep Learning for Image/Video Quality Assessment

The use of hand-crafted features for IQA has a disadvantage of being application specific and non-generic. Convolutional neural networks (CNN) can help overcome this issue by allowing an automated feature learning approach that can be extended to all kinds of distortions. CNNs

are very effective in all kinds of image processing tasks like object classification, enhancement, segmentation and retrieval. For this reason, the use of different deep neural networks (DNN) based on CNN have also been proposed for image quality assessment.

However, the use of deep learning for IQA is not a straightforward task and there are many variations on how it is applied for IQA. For instance, in [38], a blind IQA is proposed based on DNN to predict distribution of quality rating instead of the scores. They have used ResNet-101 with Huber's loss function using full size images as input. Similarly, in [39], the authors have explored VGG16, Inception-v2 and MobileNet to also predict distribution of aesthetic and technical quality ratings using Earth Mover's Distance loss function.

Ma et al. [40] have tackled IQA using a multi-task DNN using two sub-networks with shared initial layers. They have proposed to use one of the sub-networks for distortion identification and the other one for quality prediction. The output of the first sub-network is a class probability vector which is then combined with the output of the second sub-network in a weighted summation. Their approach is close to what we want to propose but to be specific, our aim is to have an opinion-unaware single network for distortion classification and ranking.

Besides these methods, some CNN based methods for IQA also make use of image patches as inputs [41, 179, 42]. Since the subjective score exists for the entire image, these methods either consider total score as an average of all patches or weighted average based on saliency or output of another sub-network. Some recent works have also proposed methods for quality based image ranking. For instance, Xu et al. [180] use machine learning based approach to predict ranking for images before evaluating the absolute score. More recently in [178], Siamese neural network has been employed to learn rankings of images. They have also used image patches as input to their neural network.

Recently, various No-Reference methods based on deep neural networks have been proposed for VQA. However, most of these works focus only on VQA for compressed videos at varying severity levels and have not been targetted and tested for other kinds of distortions. For instance, in [181], the authors have proposed a single end-to-end network for compressed videos called V-MEON (Video Multi-task End-to-end Optimized neural Network). Their network is composed of three components. The first spatiotemporal component extracts the features from the video using a deep network comprising of 3D convolutional layers. The output features from this layer are used as input by the two next components of codec classifier and quality predictor with output probability vector of codec classifier component also serving as the input to the quality predictor.

Varga [182] has proposed a simple strategy based on temporal pooling of the frame level deep features. For extracting the deep features from frames, they have fine-tuned the pre-trained Inception-V3/Inception-ResNet-v2 Convolutional Neural Networks (CNN) to classify the quality of the frame into one of the five predefined levels. In a similar approach, the authors in [62] have used Inception-ResNet-v2 pre-trained on Imagenet and extracted features from multiple layers of the model instead of one. Instead of temporal pooling they have tested two strategies. In the first strategy, they input the average feature vector to a feed forward DNN. In the second strategy, they have used deep Long Short Term Memory (LSTM) architecture. However, both these methods [182, 62] give good results for one video quality database and do not perform well on the other databases tested. Furthermore, the use of transfer learning from Imagenet is not ideal for VQA since not only the pre-training is done for object classification rather than quality level prediction but also the images in Imagenet database may not contain the kinds of distortions seen in the videos.

In [63], the authors have proposed a hybrid approach called Deep Blind VQA (DeepBVQA) whereby for spatial features they have used CNN and for temporal cues they have used hand-crafted features. For spatial feature extraction, they use a pre-trained CNN to extract the features from multiple patches for each frame. For temporal aspect, they have used simple temporal sharpness variation as the feature. Finally, before regressing to the corresponding subjective score, they aggregate all the features from frames for the video using four types of pooling namely average, variance, average upper pecentile and average lower percentile. The major drawback of their method is that they have failed to fully exploit the neural networks and instead have gone for one kind of simple hand-crafted temporal features.

Hou et al. [64] have proposed a two-stage network for NR-VQA. For the first step, they have used first 12 layers of VGG-net for extracting frame-based features. Here again the network used is pre-trained on ImageNet dataset. The features for the 8 consecutive frames from the first stage are then concatenated and given as the input to the second stage consisting of 3D CNN. However, they have also tested their results only on LIVE, CSIQ and VQEG video quality databases consisting mainly of compression and transmission based distortions. Another VQA metric proposed by [65] employs two deep networks. For the first network, they have used 3D CNN that takes small video sequences of 16 frames as input and outputs quality features of small clips. The output from the first network then act as the input to the second one consisting of LSTM model to predict the overall quality score for the video. The only drawback of their method is that their model can not be trained in an end-to-end manner since it uses two separate

models.

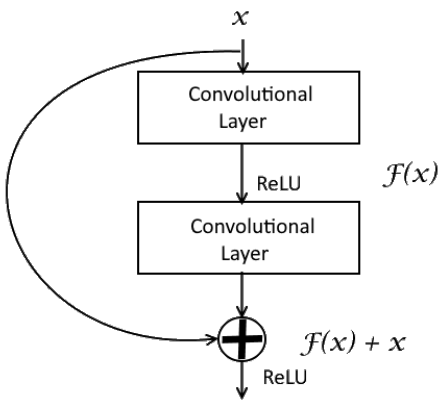


Figure 5.1: Resnet-18 basic building block

The use of Residual networks for VQA has also been recently exploited. In [66], the authors have used pre-trained Resnet-50 with global pooling as a spatial feature extractor and named it as Content-aware feature extraction. For temporal modeling, they have used Gated Recurrent Unit (GRU) network to learn long-term dependencies and for feature aggregation, followed by subjectively-inspired temporal pooling. They have designed their network specifically for in-the-wild videos and hence did not consider the aspect of distortion classification in their work. Similarly, another work for in-the-wild videos [183] has also shown promising results from ResNet. Although their proposed method VIDEVal is a type of conventional VQA and does not use neural networks, yet they have also experimented with Resnet in their comparison with other methods. However, they have used a simple strategy of average temporal pooling of output features from ResNet for predicting the scores.

5.3 Residual Network Architecture

Deeper convolutional neural networks, which simply stacked many layers that exceed a certain threshold may result in gradient vanishing problem and lead to an accuracy drop [184]. To alleviate this problem, residual network, which replaces direct stacked layers by skip connections that sum-up the output of a block of layers to its input, has been developed [184]. Figure 5.1 shows the basic building block of a residual network with skip connections whereas Figure 5.2 illustrates complete architecture of Resnet-18.

Residual network, referred to as ResNet, is considered as a class of deep neural networks which is commonly applied to image recognition as well as classification [184]. For instance, ResNet-50 is a deep network, with 50 layers and over 23 million trainable parameters, which

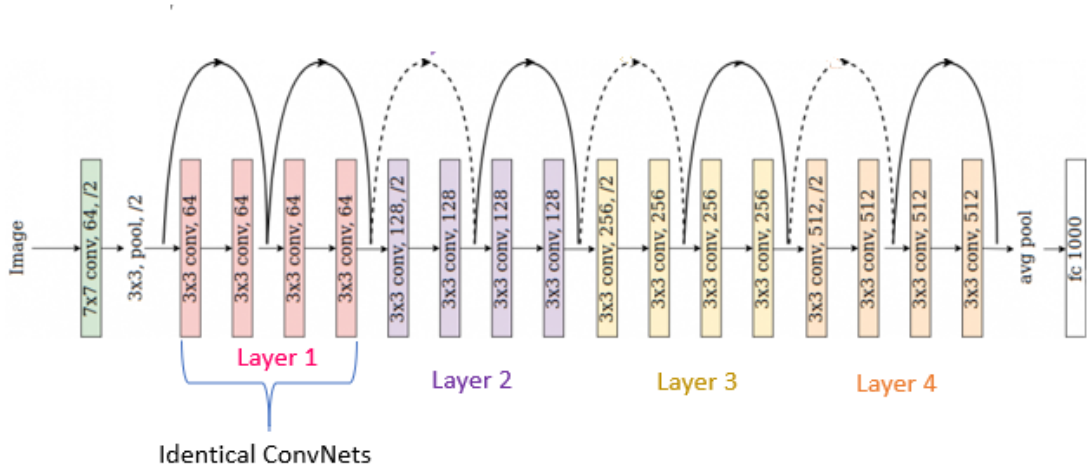


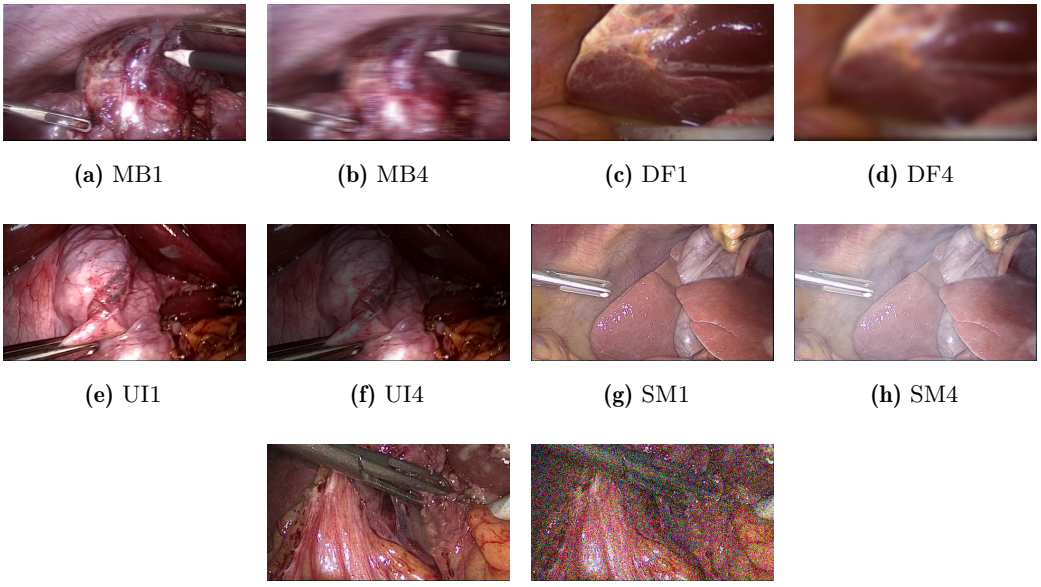
Figure 5.2: Resnet-18 architecture

has excellent generalization performance. For these reasons, we have decided to use the ResNet architecture for spatial feature modeling in our proposed VQA method. To this end, for a given ResNet architecture with a fixed number of layers, only the output fully-connected layer is replaced with a layer of size C corresponding to the number of classes in the dataset. Then, first the training of the model is done on our laparoscopic dataset for distortion classification and ranking task. This is then further extended to quality prediction task for image as well as video as explained in the next sections.

5.4 Proposed ResNet-based distortion classification and ranking

In the standard two-stage framework based image quality assessment as shown in Figure 4.1, the feature extraction step plays a crucial role for distortion classification as well as for quality score prediction. In this section, we will focus on the first stage where a ResNet-based solution is proposed to perform jointly distortion classification and ranking. For training and testing of our proposed approach, we have used the LVQ database, which was presented in the previous chapter.

Here, to be concise we denote the five distortions in LVQ database as SM for smoke, WN for additive white Gaussian noise, UI for uneven illumination, DB for blur due to defocus and MB for blur due to motion. For each distortion, the levels are labeled, from the least to the most distorted one, as hardly visible (HV), just noticeable (JN), very annoying (VA) and extremely annoying (EA). Numerically, they may be represented by numbers from 1 to 4 with 1 being the least severe and 4 being the most severe. Figure 5.3 illustrates some frames from the distorted



(a) MB1 **(b)** MB4 **(c)** DF1 **(d)** DF4

(e) UI1 **(f)** UI4 **(g)** SM1 **(h)** SM4

(i) WN1 **(j)** WN4

Figure 5.3: Illustration of the different distortion types at low and high severity levels for a given reference frame taken from the LVQ dataset.

videos from LVQ database with corresponding labels.

5.4.1 Problem formulation

In order to deal with the lack of labeled data, we propose now to aggregate the tasks of distortion classification and quality ranking as a single multi-label classification problem. More precisely, let us denote by $X^{(m)}_{d_i,l_j}$ the m-th video sequence in the dataset with distortion type d_i and severity level l_j, where

$$d_i \in \{DB, MB, WN, SM, UI\} \tag{5.1}$$

$$l_j \in \{HV, JN, VA, EA\} \tag{5.2}$$

Such multi-label classification problem can be transformed into a single-label multiclass classification task using one of the Problem Transformation methods [185]. Amongst these methods, we use the Label Powerset approach where each combination of label (d_i, l_j) can be considered as a separate class. This will result in the following set of 20 classes C:

$$C = \{(d_i, l_j)_{\substack{d_i \in \{DB,MB,WN,SM,UI\} \\ l_j \in \{HV,JN,VA,EA\}}}\}, \tag{5.3}$$

Therefore, the joint distortion classification and ranking tasks can be seen as a single-label multiclass classification problem which will now be solved using a deep learning approach.

5.4.2 ResNet-based solution

While deep convolutional neural networks have been widely investigated in the literature, it has been shown that they may lead to vanishing gradient problem and drop in accuracy performance due to the use of many stacking layers. To solve this problem, Residual Network (ResNet) has been developed by introducing skip connections [184]. For instance, the latter has been retained in different image processing tasks such as recognition [184], classification [186], etc. For this reason, we propose in this paper to resort to the ResNet architecture to solve our distortion classification and ranking problem.

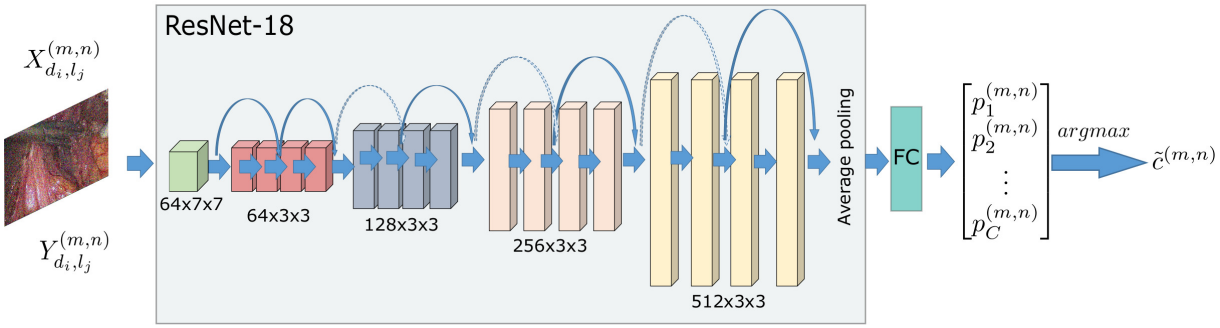


Figure 5.4: Proposed Frame-level Distortion Classification Residual Network (FDC-ResNet).

In this respect, and in order to deal with our video database, we first propose to apply separately the ResNet architecture to all the frames of the different video sequences [173]. It is important to note here that using such a frame-based approach allows us to overcome the problem of limited labeled training data for videos. The proposed solution, referred to as Frame-level Distortion Classification ResNet (FDC-ResNet), is shown in Figure 5.4. More precisely, as employed in conventional image classification problem, any given variant of ResNet architectures (i.e. ResNet-18, ResNet-34, ResNet-50, etc) is first applied to the different frames of the distorted video, which will be denoted by $X_{d_i,l_j}^{(m,n)}$ where $m \in \{1,\ldots,M\}$ represents the video sequence index in the dataset and $n \in \{1,\ldots,N\}$ refers to the frame index in that video sequence. Let us also assume that $M \times N$ training distorted frames $(X_{d_i,l_j}^{(1,1)},\ldots,X_{d_i,l_j}^{(1,N)},\ldots,X_{d_i,l_j}^{(M,1)},\ldots,X_{d_i,l_j}^{(M,N)})$ with corresponding labels $(Y_{d_i,l_j}^{(1,1)},\ldots,Y_{d_i,l_j}^{(1,N)},\ldots,Y_{d_i,l_j}^{(M,1)},\ldots,Y_{d_i,l_j}^{(M,N)})$ are available. Thus, by performing the convolution and pooling operations on the training samples in a mini-batch, a feature vector is then generated for each sample. Then, a fully connected layer, with output size equal to the number of classes C, is used with a softmax function to generate the following

vector of probability scores:

$$\mathbf{p}^{(m,n)} = [p_1^{(m,n)}, \ldots, p_C^{(m,n)}]^\top \tag{5.4}$$

where $p_c^{(m,n)}$ is the estimated probability that $X_{d_i,l_j}^{(m,n)}$ belongs to the c-th class.

This model is trained by minimizing the cross-entropy function given by

$$\mathcal{L}_\mathcal{C} = -\frac{1}{N_\mathcal{B}} \sum_{m,n} \sum_{c=1}^{C} \mathbb{1}[Y_{d_i,l_j}^{(m,n)} = c] \log(p_c^{(m,n)}) \tag{5.5}$$

where the indicator function $\mathbb{1}[Y_{d_i,l_j}^{(m,n)} = c]$ is equal to 1 when the label index $Y_{d_i,l_j}^{(m,n)}$ of the frame $X_{d_i,l_j}^{(m,n)}$ is c; otherwise it is equal to 0, and $N_\mathcal{B}$ is the mini-batch size.

Finally, the predicted class $\tilde{c}^{(m,n)}$ of the input frame $X_{d_i,l_j}^{(m,n)}$ is obtained by selecting the class yielding the maximum probability value:

$$\tilde{c}^{(m,n)} = \underset{c \in \{1,\ldots,C\}}{\operatorname{argmax}} \, (p_c^{(m,n)}) \tag{5.6}$$

Once the FDC-ResNet model is trained, it can be applied on the test video data. Indeed, for each input distorted video $X_{d_i,l_j}^{(m)}$, the trained FDC-ResNet model is applied to all the frames $X_{d_i,l_j}^{(m,n)}$ to generate their respective predicted classes $\tilde{c}^{(m,n)}$, with $n \in \{1, \ldots, N\}$. The trained network is then finally fine-tuned for only distortion classification using d_i as the set of labels. Moreover, the same training can be used for video quality prediction network, proposed in the next section.

5.5 Extension to Laparoscopic Video Quality Prediction

5.5.1 Motivation

Once the feature extraction step is done for distortion classification and ranking, a straightforward approach may consist in resorting to a regression module to map the extracted feature vector to a quality score as generally performed in two-stage framework for blind image quality assessment [32]. However, it would be more interesting to adapt the network model and make it more appropriate for quality prediction task, while taking into account the temporal effects. To this end, and similarly to the previous classification step, we first propose to perform a ResNet-based quality prediction stage on the different frames and then merge the resulting frame quality scores, based on a Fully Connected Neural Network (FCNN), to generate the final quality prediction score of a given video.

5.5.2 Modified ResNet-based solution

By following the same strategy used in the previous task, and in order to deal with the limited number of labeled training data for videos, we first propose to apply separately our ResNet-based solution on the different frames of the dataset. More precisely, the modified ResNet-based solution, referred to as Frame-level Quality Prediction Residual Network (FQP-ResNet), is shown in Figure 5.5.

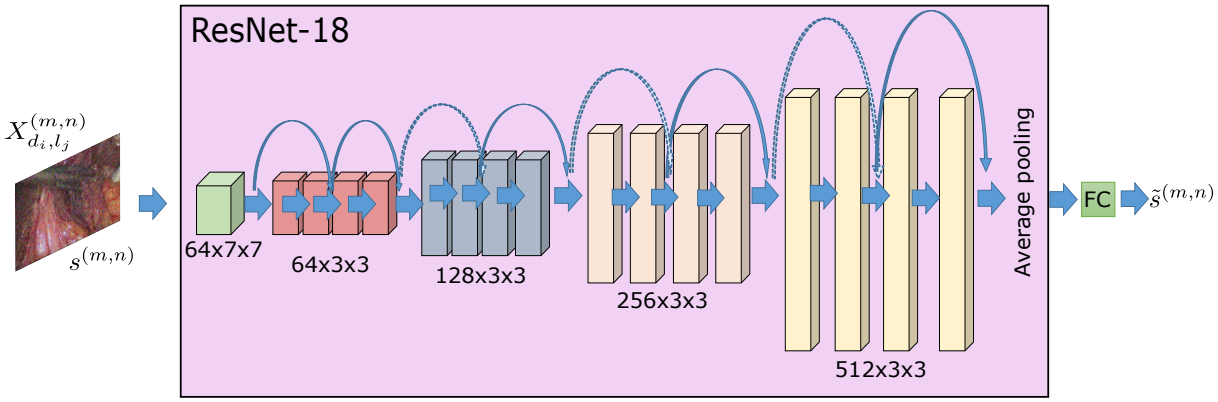


Figure 5.5: Proposed Frame-level Quality Prediction Residual Network (FQP-ResNet).

Thus, compared to the previous FDC-ResNet, and after generating the feature vector using the convolution and pooling layers, a fully connected layer of size one neuron in used in the FQP-ResNet to produce the predicted score $\tilde{s}^{(m,n)}$ associated to the n-th frame of the m-th distorted video. For training this modified ResNet-based prediction model, we propose to start from the pre-trained ResNet model used for classification (as an initialization step) and then fine-tune its different weights to adapt it to the quality prediction task. As a good quality assessment metric should be well correlated with human opinion, the loss function used to optimize the FQP-ResNet weight parameters will be defined based on the Pearson correlation coefficient often employed as a standard criterion to compare image/video quality assessment methods. To this end, we propose to minimize the following loss function:

$$\widetilde{\mathcal{L}}_{\mathcal{P}} = 1 - \frac{\sum_{m,n}\left(s^{(m,n)} - \mu_s\right)\left(\tilde{s}^{(m,n)} - \mu_{\tilde{s}}\right)}{\sqrt{\sum_{m,n}\left(s^{(m,n)} - \mu_s\right)^2 \sum_{m,n}\left(\tilde{s}^{(m,n)} - \mu_{\tilde{s}}\right)^2}} \tag{5.7}$$

where $s^{(m,n)}$ is the target subjective score associated to the n-th frame of the m-th distorted video and, μ_s and $\mu_{\tilde{s}}$ are the means of the target subjective scores and the predicted ones, respectively. It should be noted here that the subjective score, obtained during the subjective test, are provided for the overall video sequence. Hence, we will assume here that the subjective

scores of the different frames are equal to that of the corresponding video sequence. Despite its simplicity, this assumption is in some way reasonable for the following two reasons. First, in the context of laparoscopic surgery, typical video sequences do not show significant variations between the frame contents. Moreover, the temporal effect aspect will be taken into account later when merging the frame quality scores to produce a final score for the overall video.

Once the FQP-ResNet model is described, we will focus now on its application for video quality prediction. The complete architecture, referred to as Video Quality Prediction Network (VQP-Net), is illustrated in Figure 5.6.

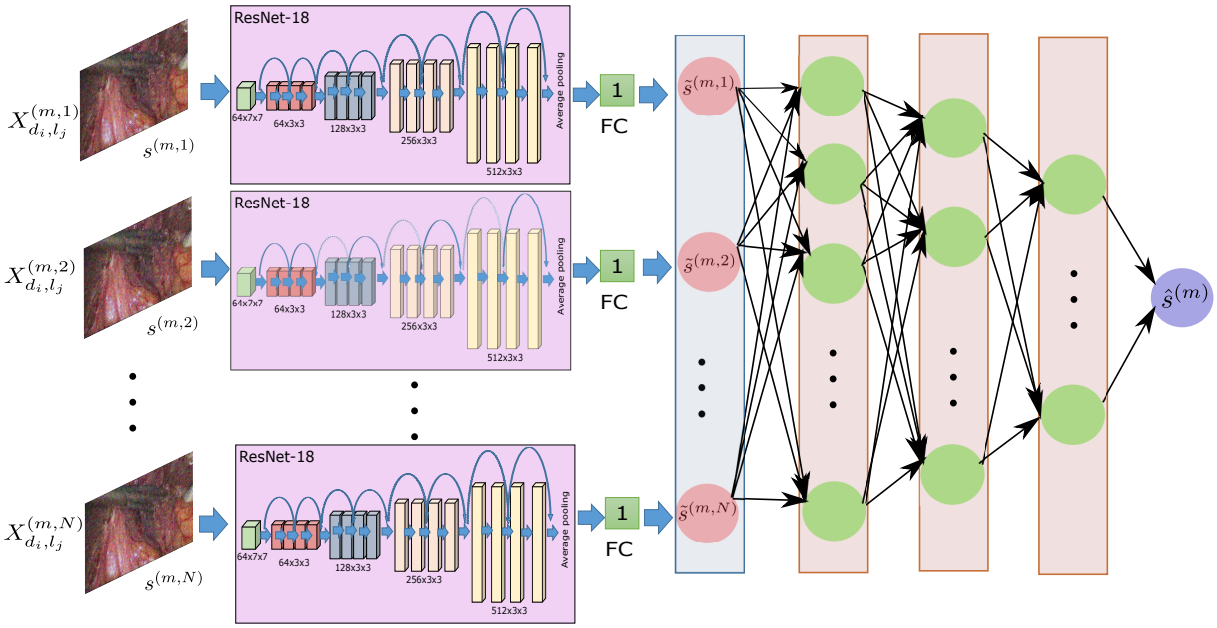


Figure 5.6: Proposed Video Quality Prediction Network (VQP-Net).

As it can be seen from Figure 5.6, for any input distorted video sequence $X_{d_i,l_j}^{(m)}$, the previous FQP-ResNet is first separately applied to the different frames $X_{d_i,l_j}^{(m,n)}$ resulting in the following subjective quality score vector $\tilde{\mathbf{s}}^{(m)}$:

$$\tilde{\mathbf{s}}^{(m)} = [\tilde{s}^{(m,1)}, \ldots, \tilde{s}^{(m,N)}]^\top \tag{5.8}$$

Then, in order to take into account the temporal effect, an additional FCNN is incorporated to combine the different frame quality scores and produce a final video quality score. More precisely, the subjective quality score vector $\tilde{\mathbf{s}}^{(m)}$ is associated to the input layer of this FCNN. Then, H hidden layers are used where their neuron values are computed from the previous ones based on a linear combination (with bias) followed by a nonlinear activation function. Finally, an output layer with a single neuron is employed yielding the computation of the predicted quality score

$\hat{s}^{(m)}$ associated to the input distorted video sequence $X^{(m)}_{d_i,l_j}$.
For the training of this proposed VQP-Net model, two approaches will now be addressed.

5.5.2.1 Transfer learning approach

One straightforward approach would consist in following a process similar to the transfer learning which is based on the use of a pre-trained model. Thus, the previous pre-trained FQP-ResNet model will be firstly used by the first block of the overall architecture (see Figure 5.6) to generate the frame quality scores. Then, our training strategy will focus only the learning of the FCNN weight parameters. This will be achieved by maximizing the correlation between the target subjective score $s^{(m)}$ and the predicted one $\hat{s}^{(m)}$. Therefore, the FCNN weight parameters are updated by minimizing the following Pearson correlation coefficient based loss function:

$$\mathcal{L}_{\mathcal{P}} = 1 - \frac{\sum_m \left(s^{(m)} - \mu_s\right)\left(\hat{s}^{(m)} - \mu_{\hat{s}}\right)}{\sqrt{\sum_m \left(s^{(m)} - \mu_s\right)^2 \sum_m \left(\hat{s}^{(m)} - \mu_{\hat{s}}\right)^2}} \tag{5.9}$$

Once the FCNN model is trained, the obtained learned parameters as well as those of the pre-trained ResNet one are employed by the complete VQP-Net architecture for computing the prediction quality score of each input test video sequence.

5.5.2.2 End-to-end learning approach

While the previous approach presents the advantage of reducing the complexity of the training phase, a more interesting approach should consist in resorting to an end-to-end learning approach. More precisely, instead of using the pre-trained FQP-ResNet model (while keeping it fixed) in the first block of the VQP-Net architecture, we propose here to update the weight parameters of the FQP-ResNet model during the training phase. Therefore, the parameters of both the FQP-ResNet model (i.e first block) and the FCNN one (i.e second block) of our VQP-Net architecture are simultaneously optimized by using the same loss function defined in (5.9). The final learned parameters of the whole VQP-Net model are then used to predict the quality scores of the test distorted videos.

To show the interest of this end-to-end learning approach compared to the transfer learning one, Figure 5.7 illustrates the loss functions (given by (5.9)) evaluated on the validation dataset with respect to the number of epochs. Thus, it can be observed the end-to-end training approach outperforms the transfer learning one in terms of loss function values. Moreover, the convergence of the end-to-end training approach is much faster than that of the transfer learning technique.

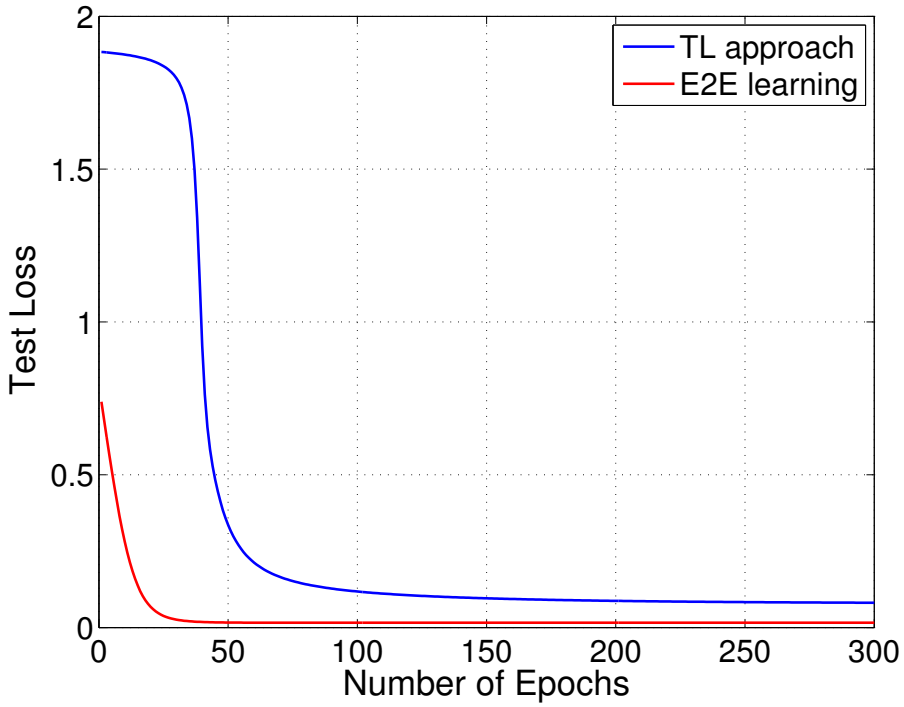


Figure 5.7: Loss function evolution with the number of epochs for the transfer learning and end-to-end learning approaches.

5.6 Experimental Results

In this section, intensive experiments have been conducted to evaluate the performance of the proposed video distortion classification and quality prediction methods. In the following, we will define the experimental settings used in our simulations, present the comparison methods and finally discuss the obtained results.

5.6.1 Experimental settings

The proposed ResNet based methods were tested on the LVQ dataset presented in Chapter 4. For the task related to the distortion classification and ranking (FDC-ResNet), the original training dataset has been extended by creating more distorted videos using the Cholec80 dataset [166] and following the methodology explained in [155]. Furthermore, a frame-level data augmentation step is also performed by applying random cropping and random horizontal flipping to the original frames. It is important to note here that we did not carry out any subjective test for these new videos since the subjective scores are not used for training the FDC-ResNet. Indeed, the latter only requires the information about the rank and distortion type. After that, for the quality prediction task (FQP-ResNet), the obtained FDC-ResNet model is fine tuned and trained on

the original training dataset with available subjective scores. Regarding the VQP-Net task, the FCNN employed in the second block of the overall architecture is implemented using three hidden layers (i.e $H = 3$) and the log softmax function as an activation function. Note that 80% of the original LVQ database is used for training and 20% for testing.

The implementation of the proposed network was done using Pytorch and the network was run on a Windows system with Nvidia Quadro RTX-6000 GPU and 32 GB RAM. For the training of FDC-ResNet, we used the Adam optimizer with a learning rate of 0.01. The learning rate was dynamically reduced if the loss value did not change for two consecutive iterations. For training of VQP-Net, a lower learning rate of 0.00001 was used with the Adam optimizer.

5.6.2 Comparison methods

To demonstrate the effectiveness of the proposed video objective quality prediction method, a comparative study with some state-of-the-art VQA metrics is conducted. A first way to evaluate the quality of videos is to consider state-of-the-art IQA metrics designed for images, such as SSIM [23], VIF [187], BRISQUE [35] and NIQE [172], apply them frame by frame and use a temporal pooling model to derive VQA measure of the whole video stream. This simple approach was used as the first intuitive solution for estimating the global video quality. For these metrics, we have applied four different temporal pooling techniques for combining the predicted scores from all the frames namely the average pooling, geometric mean pooling, harmonic mean pooling and median pooling. We also considered quality metrics designed specifically for video and which integrate, in one way or another, the temporal aspect according to some pooling models in an explicit or implicit way. These VQA measures can be classified into conventional and deep learning (DL) based metrics. The conventional metrics used for comparison are V-BLIINDS [60], VIIDEO [36] and TLVQM [188]. For DL based methods, we tested four recent approaches namely Inceptionv3-FT [182], VSFA [66], CNN-LSTM TLVQM [189] and CNN-SVR TLVQM [189]. It should be noted here that these methods have been tested using the source codes provided by the authors except for the Inceptionv3-FT method which is tested using the corrected implementation [62]. For a fair comparison, we have retrained all these deep learning based methods on our LVQ dataset.

The performance of the different video quality prediction methods is evaluated in terms of various standard criteria measuring the correlation between the predicted scores and the subjective ones. More precisely, we have computed the Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC) and Kendall Rank-Order

Methods	SROCC
FDC-ResNet	**0.69**
RankIQA [178]	0.57

Table 5.1: SROCC of ranking methods with laparoscopic dataset

Proposed method with	Accuracy
ResNet-18 with augmented data (FDC-ResNet)	97.8%
ResNet-18	83.3%
ResNet-34	84.7%
ResNet-50	**87.3**%

Table 5.2: Classification accuracy of the proposed method with different ResNet models for our laparoscopic dataset.

Correlation Coefficient (KROCC) after performing five-parameter logistic regression for the predicted scores.

5.6.3 Results and discussion

5.6.3.1 Distortion ranking performance

: To the best of our knowledge, our methodology is the first one which aims to perform distortion classification as well as ranking simultaneously. For this reason, we will propose to compare our approach separately to classification methods and ranking methods. Indeed, the results of quality ranking from our method are firstly compared to a recent deep learning based method which employs a Siamese network for ranking before fine-tuning its single trained branch to obtain quality scores [178]. The latter is designated by RankIQA. For our comparison, we have only used the ranking part of their network with Hinge loss and have retrained it on LVQ database. In order to evaluate these ranking methods, we have used Spearman Rank-Order Correlation Coefficient (SROCC). Table 5.1 provides the SROCC values for the RankIQA method as well as the proposed one using ResNet-18. The obtained results show that our approach yields better ranking performance compared to the recent deep learning-based RankIQA method.

Table 5.3: Comparison of distortion classification accuracy.

Method	Classification Accuracy
FDC-ResNet	97.8%
BRISQUE	94.2%
BIQI	95.0%
BLIINDS-II	95.6%

5.6.3.2 Distortion classification performance

For evaluating performance of distortion classification, first we have evaluated the classification performance of our proposed architecture and have compared it two other deeper variants with more layers as well as with and without data augmentation. Table 5.2 shows the mean classification accuracy, for ResNet-18 with and without data augmentation, ResNet-34 and ResNet-50. Thus, it can be observed that the used ResNet-18 with augmented data leads to good accuracy results (around 97.8%). Moreover, by using deeper ResNet architectures, the accuracy is also improved even without data augmentation and reaches 87.3% with ResNet-50 compared to only 83.3% for ResNet-18.

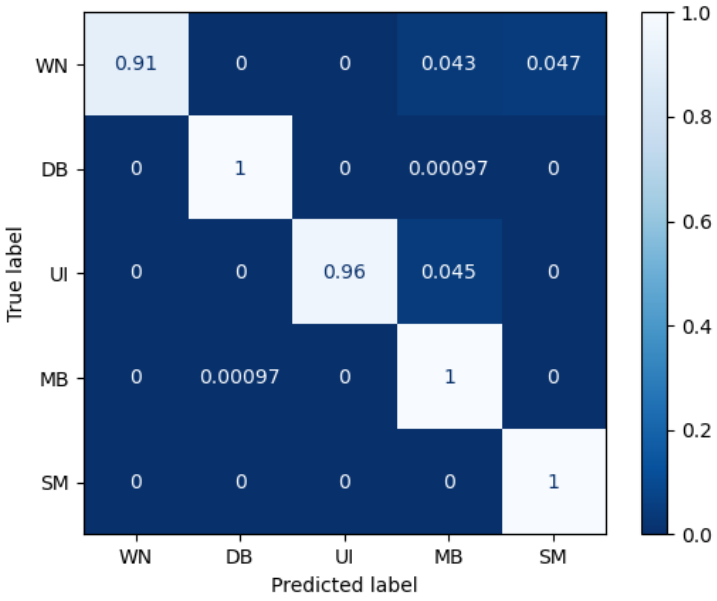


Figure 5.8: Confusion Matrix for Proposed FDC-ResNet.

Besides comparing different ResNet models, we have also compared the performance of our method with some existing classification methods. In this regards, we have selected some of the conventional IQA metrics based on the two-stage network which also perform the classification namely BRISQUE, BIQI and BLIINDS-II. Table 5.3 shows the accuracy results of these methods. Thus, it can be noticed that our proposed method outperforms the conventional ones. Furthermore, Figure 5.8 illustrates well, through the confusion matrix, the superiority of the proposed distortion classification method in terms of prediction error and more particularly in the case of smoke, blur due to motion and defocus blur.

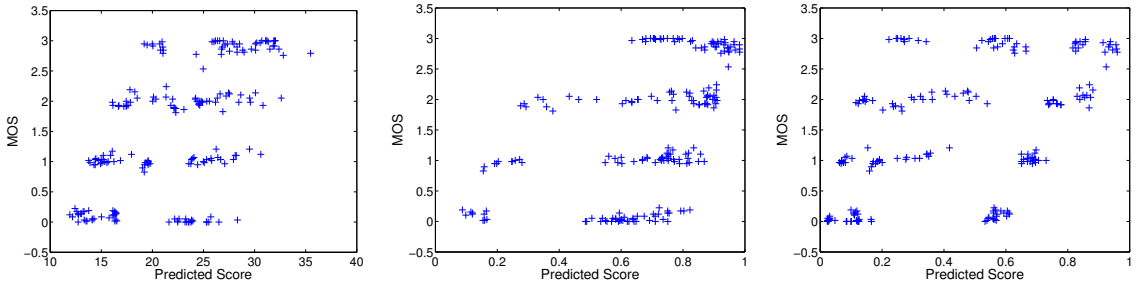
5.6.3.3 Video quality prediction performance

Table 5.4 shows the results of our proposed methods as well as the state-of-the-art ones for the LVQ database in terms of PLCC, SROCC and KROCC. Note that the category of metric to which belongs each metric (conventional or deep learning (DL) one) as well as its type (full-reference (FR) or no-reference (NR) one) are also provided. From the table, it can be firstly observed that the IQA based methods perform poorly in terms of PLCC, SROCC and KROCC, even when using temporal pooling strategies other than the average pooling one. Among these IQA-based video quality assessment methods, the highest correlation values are obtained using the PSNR metric with median pooling. Secondly, regarding the NR-VQA methods, except for the VIIDEO metric, TLVQM and V-BLIINDS ones outperform all the conventional IQA-based video quality metrics. Indeed, V-BLIINDS leads to the best results compared to the other conventional IQA-based metrics and VQA ones. The obtained PLCC, SROCC and KROCC values are 0.8328, 0.8242 and 0.6317, respectively. Finally, concerning the DL based VQA methods, one can observe that the performance of the CNN-TLVQP method depends on the temporal pooling strategy. Indeed, while the Support Vector Regression (SVR) based temporal pooling stage is less performant compared to some conventional methods, the Long Short-Term Memory (LSTM) based temporal stage improves significantly the results yielding a gain of around 0.2 in all the correlation values. Further improvements are achieved by the Inception-v3 method, and more specifically, the VSFA method. For the proposed VQP-Net method, it can be firstly noticed that its first version based on the transfer learning (TL), outperforms the state-of-the-art deep learning ones except the VSFA method. Most importantly, the end-to-end learned version (E2E) allows us to improve the VSFA method while achieving a gain of around 2.5 % in terms of PLCC and 1.5 % in terms of SROCC and KROCC.

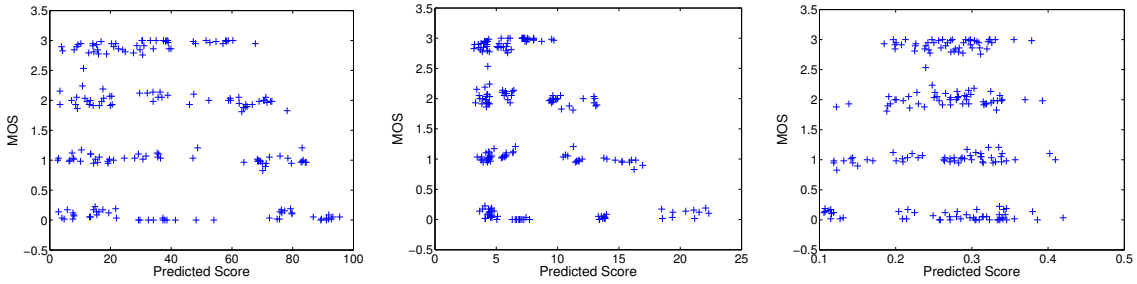
In order to have a better idea of the prediction accuracy for these VQA methods, Figure 5.9 shows the scatter plots of subjective scores (MOS) against the predicted scores. For the

Table 5.4: PLCC, SROCC and KROCC for the scores of the different video quality prediction methods.

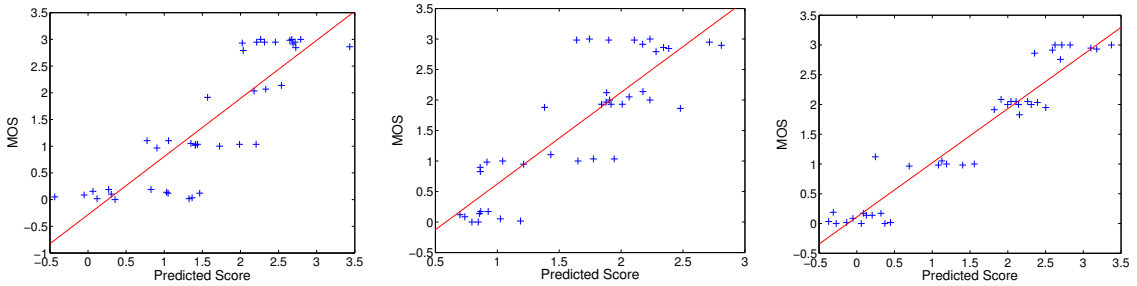
Metric	Category	Type	Temporal Model	PLCC	SROCC	KROCC
PSNR	Conventional	FR-IQA	Average Pooling	0.6973	0.6996	0.5030
			Geometric Mean Pooling	0.6900	0.6949	0.4965
			Harmonic Mean Pooling	0.6778	0.6829	0.4875
			Median Pooling	0.7131	0.7097	0.5133
SSIM [23]	Conventional	FR-IQA	Average Pooling	0.6123	0.5914	0.4187
			Geometric Mean Pooling	0.6034	0.5812	0.4123
			Harmonic Mean Pooling	0.5902	0.5563	0.3980
			Median Pooling	0.6157	0.5945	0.4183
VIF [187]	Conventional	FR-IQA	Average Pooling	0.6267	0.6228	0.4537
			Geometric Mean Pooling	0.6158	0.6307	0.4503
			Harmonic Mean Pooling	0.5834	0.6081	0.4397
			Median Pooling	0.6211	0.6192	0.4504
BRISQUE [35]	Conventional	NR-IQA	Average Pooling	0.4593	0.4304	0.3108
			Geometric Mean Pooling	0.4609	0.4300	0.3114
			Harmonic Mean Pooling	0.4722	0.4360	0.3199
			Median Pooling	0.4639	0.4310	0.3119
NIQE [172]	Conventional	NR-IQA	Average Pooling	0.4242	0.3731	0.3555
			Geometric Mean Pooling	0.4535	0.4958	0.3594
			Harmonic Mean Pooling	0.4402	0.4856	0.3641
			Median Pooling	0.4583	0.4994	0.3633
VIIDEO [36]	Conventional	NR-VQA	Temporal Features	0.3842	0.3416	0.2313
V-BLIINDS [60]	Conventional	NR-VQA	Temporal Features	0.8328	0.8242	0.6317
TLVQM [188]	Conventional	NR-VQA	Average Pooling	0.7681	0.6892	0.5132
CNN-SVR TLVQM [189]	DL	NR-VQA	SVR	0.5826	0.5888	0.4088
CNN-LSTM TLVQM [189]	DL	NR-VQA	LSTM	0.8006	0.7829	0.5828
Inceptionv3-FT [182] [190]	DL	NR-VQA	Average Pooling	0.8550	0.7978	0.6202
VSFA [66]	DL	NR-VQA	GRU + Subjectively-inspired	0.9647	0.9247	0.7629
Proposed VQP-Net (TL)	DL	NR-VQA	Fully-connected Network	0.8992	0.8434	0.6494
Proposed VQP-Net (E2E)	DL	NR-VQA	Fully-connected Network	**0.9899**	**0.9388**	**0.7739**



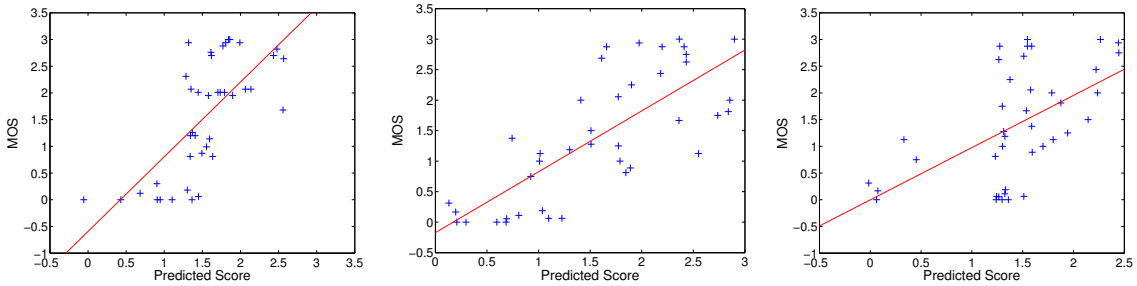
(a) PSNR **(b)** SSIM **(c)** VIF



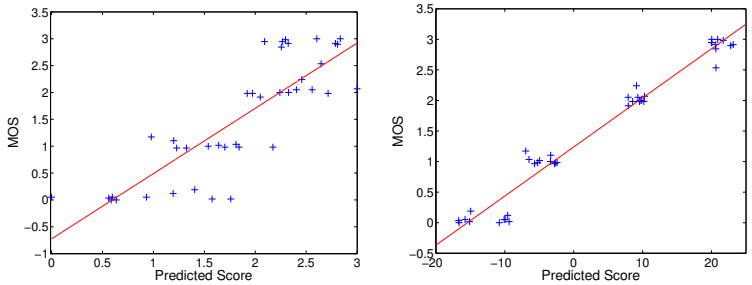
(d) BRISQUE **(e)** NIQE **(f)** VIIDEO



(g) V-BLIINDS **(h)** Inceptionv3-FT **(i)** VSFA



(j) TLVQM **(k)** CNN-LSTM TLVQM **(l)** CNN-SVR TLVQM



(m) Proposed VQP-Net (TL) **(n)** Proposed VQP-Net (E2E)

Figure 5.9: Subjective (MOS) vs predicted score plots for different VQA metrics.

conventional IQA metrics and the VIIDEO one (Figure 5.9a - 5.9f), the plots are obtained by using the scores of all the LVQ database in the plots. For the remaining plots, only the scores for the test data have been considered. Ideally, for a good VQA method, the scatter plot should show good linearity, tight clustering and a relatively uniform density along both axes. From the figure, we can see that with IQA metrics and VIIDEO (Figure 5.9a - 5.9f), there is no clear relationship between the two data and they are widely spread throughout. In the other plots (Figure 5.9g - 5.9n), the data has been linearly fitted with a straight line. If the points on each plot are closer to the fitted lines, it implies a higher consistency between the subjective and predicted scores. Hence, we can clearly see the performance of the different metrics based on the spread of the data points around the straight line. For instance, for TLVQM based methods (Figure 5.9j - 5.9l), the data is more spread as compared to the other methods, especially for the CNN-SVR TLVQM metric where the predicted scores are significantly biased towards higher values. On the other hand, for V-BLIINDS and Inceptionv3-FT (Figure 5.9g - 5.9h), we can observe that data points are closer and more uniformly distributed as compared to the previously mentioned methods. Similar behavior is also observed with our proposed VQP-Net (TL) method with transfer learning approach. However, with VSFA method (Figure 5.9i), we see even much better clustering than our proposed transfer learning based approach, even though some of the data points are still away from the fitted line. Undoubtedly, from among all the plots, our proposed VQP-Net (E2E) with the end-to-end learning approach gives the best linearity, clustering and uniform distribution along the two axes as shown in Figure 5.9n. This further consolidates the observations made based on the correlation coefficient values.

5.6.3.4 Comparison with different temporal pooling approaches

In order to show the importance of the additional FCNN model to combine the different frame quality scores, we have performed an ablation study whereby this temporal pooling stage is achieved using other approaches based on neural networks as well as conventional operations. In this respect, we have chosen three different networks which have been considered in previous deep learning based VQA methods [66, 189]. These networks are Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN). For fair comparison, the overall architecture composed of ResNet followed by the chosen network is trained in an end-to-end manner. Regarding the conventional temporal pooling methods, we have used arithmetic mean, geometric mean, harmonic mean and median pooling approaches. Note that the latter are simply applied to the quality scores $\tilde{s}^{(m,n)}$ obtained with each frame using the FQP-ResNet model.

The results of these different temporal pooling approaches are shown in Table 5.5. Thus, in case of conventional temporal pooling methods, it can be firstly noticed that both arithmetic mean pooling and median pooling give good results which are close to those obtained using an FCNN with a transfer learning strategy. However, the results of these simple temporal pooling approaches are much less performant than the end-to-end learning version of the FCNN-based approach. Moreover, it can be also observed that the retained FCNN model outperforms significantly the other neural network ones.

Table 5.5: Comparison of different temporal pooling methods for combining the frame quality scores.

Methods	PLCC	SROCC	KROCC
Proposed FCNN (E2E)	**0.9899**	**0.9388**	**0.7739**
Proposed FCNN (TL)	0.8992	0.8434	0.6494
RNN (E2E)	0.7768	0.7039	0.5276
GRU (E2E)	0.6938	0.6331	0.4550
LSTM (E2E)	0.5434	0.4880	0.3461
Arithmetic Mean Pooling	0.9019	0.8459	0.6494
Geometric Mean Pooling	0.7728	0.7112	0.4939
Harmonic Mean Pooling	0.7803	0.7190	0.5276
Median Pooling	0.9033	0.8456	0.6520

5.7 Conclusions and Perspectives

In this chapter, we presented a novel method for quality assessment of laparoscopic videos using a combination of residual network and a three-layered fully-connected network. Residual network is used to capture spatial features from each frame whereas fully-connected network is used to combine the frame-level scores for prediction of final video quality score. Moreover, our proposed approach also classifies the distortion affecting each video by using a combination of ResNet and max pooling, which is also an essential step for quality monitoring applications. We have tested our method on LVQ dataset and have shown that it outperforms all the existing conventional and deep learning based VQA methods. However, some of the important aspects which still need to considered is to evaluate the performance on real distortions and mixtures of distortions. For this, there is a need to have more rich large databases for quality assessment.

CHAPTER 6

A Multi-Criteria Contrast Enhancement Evaluation Measure using Wavelet Decomposition

Abstract

An effective contrast enhancement method should not only improve the perceptual quality of an image but should also avoid introducing artifacts or affecting naturalness of images. This makes Contrast Enhancement Evaluation (CEE) a challenging task in the sense that both the improvement in image quality and unwanted side-effects need to be accounted for. Currently, there is no single CEE metric that works well for all kinds of enhancement criteria. In this chapter, we propose a new Multi-Criteria CEE (MCCEE) measure which combines different metrics effectively to give a single quality score. In order to fully exploit the potential of these metrics, we have further proposed to apply them on the decomposed image using wavelet transform. This new metric has been tested on medical Computed Tomography (CT) images as well as on two natural image contrast enhancement databases. The results show a substantial improvement as compared to the existing evaluation metrics. [1]

[1] [191] **Khan, Z. A.**, Beghdadi A., Alaya Cheikh, F., Kaaniche, M. and Qureshi, M. A. "A Multi-Criteria Contrast Enhancement Evaluation Measure using Wavelet Decomposition". IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP) 2020.

6.1 Introduction

Contrast enhancement of images is an important processing step in medical imaging [192] as well as in other applications such as remote sensing [193] and underwater imaging [194]. A plethora of CE methods has been proposed in the literature [195]. However, besides changing contrast, most of the time these methods also produce undesirable effects such as halo-effect, unnaturalness, color bleeding or over-enhancement. In order to judge the quality of the results from these methods, there is therefore a dire need of an evaluation method that is not only efficient but also consistent with different criteria related to the given applications.

Conventional Image Quality Metrics (IQMs) aim to predict the image quality in terms of degradation with respect to a pristine image. However, CEE is a completely different task where there is no defined pristine image and the judgment is strongly subjective in the sense that there is no objective measure on how to quantify some attributes such as pleasantness and naturalness that are strongly affected by the enhancement process. Despite the challenging nature of CEE, some metrics have been proposed in the literature which try to judge the quality of enhancement objectively. While some of these metrics perform well for limited applications or images, yet none of them is robust and consistent throughout all the available CEE databases [196, 197].

The target of most of the previous works on CEE was to estimate only the improvement in contrast. Examples of such metrics include Absolute Measure of Enhancement (AME), Measure of Enhancement (EME) [198], Image Enhancement Measure (IEM) [88] and Radial Spectral Energy (RSE) [87]. Some of the other metrics have been focused only on checking preservation of some quantity in an image like Absolute Mean Brightness Error (AMBE) [79] for brightness, Edge Content (EC) [85] for edges, and mutual information based measure [199, 82]. Besides, more recently some focus has been given on over-enhancement measures like Lightness Order Error (LOE) [91], Lightness Order Measure (LOM) [92] and Struture Measure Operator (SMO) [90]. However, in reality, different contrast enhancement methods impact these quantities in different manners and an effective CEE should be able to capture all these aspects. This is achieved using a new approach where such aspects are expressed by combining multiple criteria to derive a single CEE metric.

In this new scheme, the different metrics are combined in order to exploit advantages of each metric. For instance, in [78], authors have shown how combining different metrics can improve upon their individual performances. They have especially shown the effectiveness of LOE metric [91] when combined with other CEE metrics. LOE measure aims to assess the preservation of naturalness in a contrast-enhanced image by using relative lightness order difference between the

original and the enhanced images. The lightness of each pixel is first compared to that of every other pixel to see if it is greater or less than the latter. If this lightness order is different in original and enhanced images, it signifies a change in lightness order and is recorded as an increment to the output pixel value. A larger LOE value implies a poorer preservation of naturalness in image.

Shokrollahi *et al.* [93] have also proposed in their work a similar scheme and introduced a metric for CEE called the contrast-changed image quality (CCIQ). The metric combines contrast-measuring quantities like edge-based contrast criterion (ECC) and entropy with over-enhancement detecting measures like correlation coefficient and AMBE. To evaluate the optimal weight values of these four quantities, they have employed Particle Swarm Optimization (PSO) algorithm.

The main contributions of this work are the following:

- designing a new multi-criteria scheme to compute a single CEE metric in a collaborative way so as to capture various features of the processed image,
- constructive use of selective wavelet subbands so as to capture efficiently the different aspects of CE side-effects,
- illustrate the effectiveness of the proposed metric for medical images using a post-processing task-based evaluation approach,
- validation of the new CEE metric on two natural image contrast datasets to show consistency with subjective evaluation results. .

The remainder of the chapter is organized as follows. Section 6.2 highlights different criteria needed for an effective CEE. This is followed by Section 6.3, where we describe the motivation for using image decomposition and give explanation of the steerable pyramid wavelet decomposition method used for our metric. In Section 6.4 we present our proposed CEE metric and provide a detailed explanation of our task-based CEE approach for CT images. Finally, in Section 6.5, the results of our experiments are discussed and followed by conclusions and perspectives in Section 6.6.

6.2 Contrast Enhancement Evaluation Criteria

The evaluation of CE is indeed a challenging problem due to the involvement of many factors that are mostly subjective and difficult to model. The main factor that is highly related to the quality of the enhanced image and the visibility of details is the local contrast. Various metrics for CEE have been proposed in the literature [78]. However, all of them have some limitations

as they only cater to only one aspect of CEE like detection of over-enhancement or contrast improvement. In the following, we discuss the most important aspects that should be taken into account in the design of CEE measure. We also discuss some of the most representative CEE metrics and highlight their limitations.

6.2.1 Contrast Improvement

To evaluate a contrast enhancement algorithm, the most important performance indicator is the improvement of contrast. One of earliest measures of contrast was the Weber contrast [81] which is used to measure the local contrast of a small target of uniform luminance against a uniform background. However, it is effective only for a limited context of applications and could not be extended to complex images. A much more effective metric based on Weber-Fechner's law for evaluating the contrast in complex images was proposed in [198]. It is called Measure of enhancement by entropy (EMEE). For an enhanced image I_e divided into $k_1 \times k_2$ blocks $B(i,j)$ with center (i,j), it is defined as

$$EMEE = \frac{1}{k_1 \times k_2} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \alpha \left(\frac{I_{e(ij)}^{max}}{I_{e(ij)}^{min} + C} \right)^{\alpha} \ln \frac{I_{e(ij)}^{max}}{I_{e(ij)}^{min} + C} \tag{6.1}$$

the notations max and min correspond to the maximum and minimum intensity values of the image block $I_{e(ij)}$, C is a constant to avoid the division by zero and α is an exponent that controls the enhancement effect [198].

6.2.2 Brightness Preservation

A contrast enhancement algorithm ideally should not change the overall brightness of the image. The metric called Absolute Mean Brightness Error (AMBE) evaluates the change in brightness as the difference in mean intensity values of the original image I_o and the enhanced image I_e.

$$AMBE = |E(I_o) - E(I_e)| \tag{6.2}$$

where $E(.)$ represents the mean value of the underlying image.

A lower value of $AMBE$ corresponds to a better preservation of brightness. However, using $AMBE$ alone to evaluate enhancement quality is not sufficient as it does not take into account other important factors like preservation of naturalness and edginess information in an image. Figure 6.1 further illustrates this point where $AMBE$ and MOS values for two enhanced images from CEED2016 database [196] are shown. It can be clearly seen how $AMBE$ is unable to

(a) Original **(b)** MOS=51,AMBE=2.23 **(c)** MOS=77,AMBE=46.9

Figure 6.1: Comparison of AMBE and MOS for an example of image taken from CEED2016

correlate with the subjective evaluation score in this case. We can observe from the Figure 6.1 that although the MOS value for Figure 6.1c is higher than that of Figure 6.1b, yet AMBE value is also higher for Figure 6.1c, suggesting a poorer brightness preservation as compared to Figure 6.1b.

6.2.3 Structure Preservation

Some enhancement algorithms like histogram equalization based methods, may also introduce artifacts in the images resulting in structural changes like those in edge and texture information [90]. It is important to preserve this information from the original image during contrast enhancement thereby preventing unnaturalness in the resulting image. Structure Measure Operator (SMO) [90] is an over-enhancement metric which tries to capture this structural change. Unlike quality metric Structural Similarity Index ($SSIM$) [23], it is insensitive to changes in contrast and hence is most suitable for measuring structure changes during contrast enhancement. SMO is defined in terms of the difference in pixel non-homogeneity value of the original image NHO^o_{mn} and that of the enhanced image NHO^e_{mn}:

$$SMO = \frac{1}{W \times H} \sum_{m=1}^{W} \sum_{n=1}^{H} \frac{|NHO^o_{mn} - NHO^e_{mn}|}{NHO^o_{mn}} \tag{6.3}$$

where NHO_{mn} is a product of three quantities of edge value, entropy and standard deviation in a $d \times d$ window and W and H are the width and height of the image respectively. A lower value of SMO corresponds to better preservation of structure. Figure 6.2 further shows how SMO is able to capture the over-enhancement caused by structural information changes.

(a) Original **(b)** MOS=73.5,SMO=2.13, AMBE=41.16 **(c)** MOS=0,SMO=134.3, AMBE=24.02

Figure 6.2: SMO values for an example of image with visible over-enhancement from CEED2016

6.2.4 Lightness Order Preservation

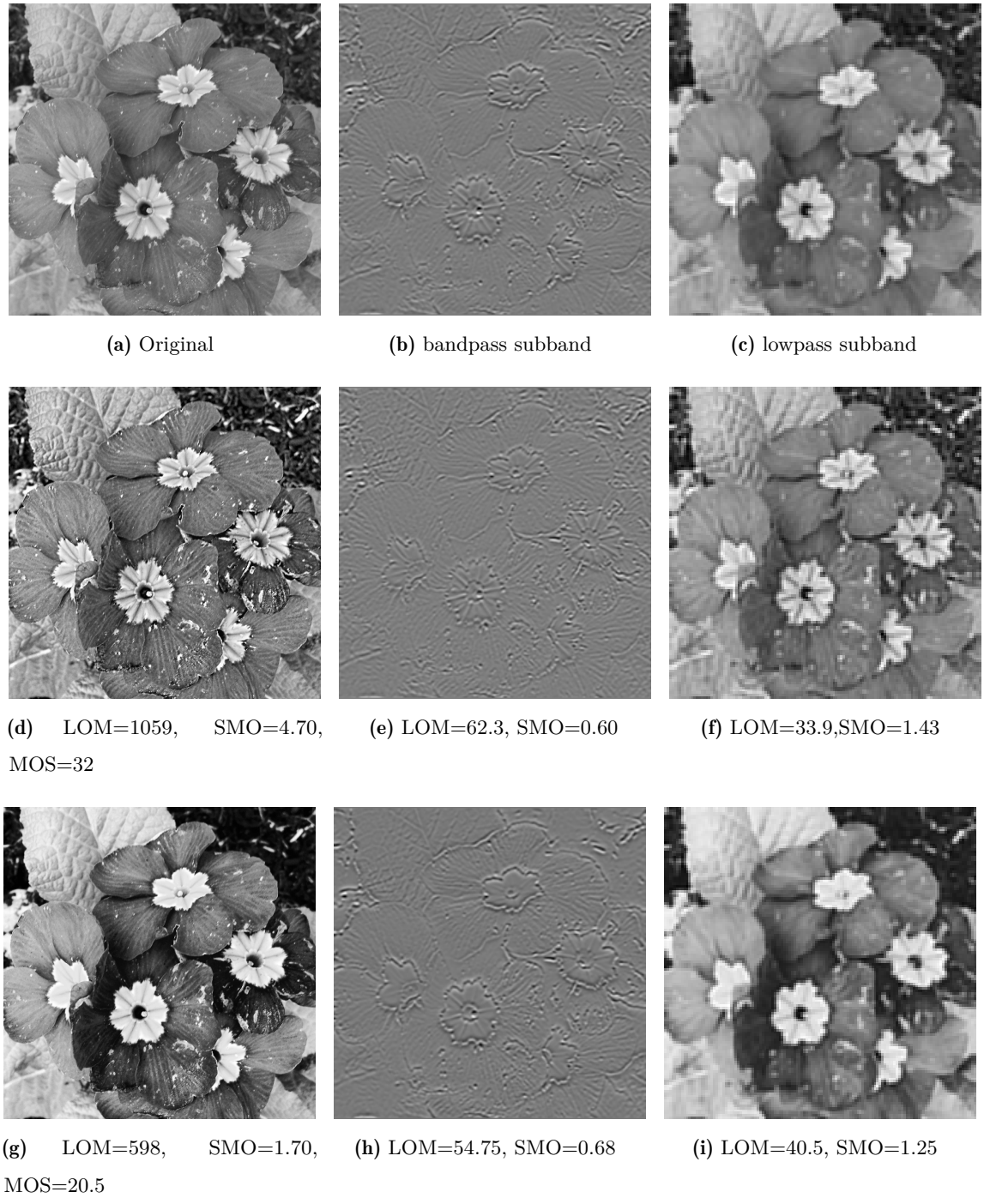
This is another important aspect of measuring preservation of naturalness for a contrast enhancement method. Lightness order is a measure of the order of lightness differences between each pair of pixels. In case of over-enhancement, this lightness order may be altered. So, evaluating the error in this lightness order can be a good measure of naturalness preservation. This has been proposed in [91] as a metric called Lightness Order Error (LOE). However, as it has been shown in [92], LOE is dependent on image content and also does not take into account *local* relative lightness order. To overcome these drawbacks, Bai *et al.* [92] have proposed another metric called the Lightness Order Measure (LOM). It is an over-enhancement measure that is based on local inversion of lightness order. It first filters the original and enhanced images with a local mean filter of a window size of 31×31. The non-filtered images are then subtracted from the filtered images to produce d_o and d_e for original and enhanced images, respectively. Finally, their difference is used to evaluate LOM as follows:

$$LOM = \frac{1}{W \times H} \sum_{m=1}^{W} \sum_{n=1}^{H} |(d_e(m,n) - d_o(m,n)).\frac{sign(d_e(m,n)) - sign(d_o(m,n))}{2}| \tag{6.4}$$

where $d_o(m,n)$ is the difference between the pixel values (m,n) of the original and its local mean-filtered version and similarly $d_e(m,n))$ is the corresponding difference for the enhanced image [92]. Here $sign(.)$ is the signum function.

6.3 Wavelet Decomposition for CEE

Existing CEE metrics consider only the original image representation for evaluating the quality of enhancement. However, very often it is useful to decompose the image into other forms like



(a) Original **(b)** bandpass subband **(c)** lowpass subband

(d) LOM=1059, SMO=4.70, MOS=32 **(e)** LOM=62.3, SMO=0.60 **(f)** LOM=33.9,SMO=1.43

(g) LOM=598, SMO=1.70, MOS=20.5 **(h)** LOM=54.75, SMO=0.68 **(i)** LOM=40.5, SMO=1.25

Figure 6.3: Over-enhancement measures LOM and SMO with input (left) and decomposed images for (a)-(c)original image (d)-(f)enhanced image and (g)-(i) over-enhanced image

low-pass, high-pass and bandpass components to gain an insight on the effects of CE method applied. For instance, any deterioration in image structure is more detectable in the bandpass subband whereas changes to the global image properties are more likely to be detected in lowpass subband.

To further illustrate this point, Figure 6.3 shows three images from CEED 2016 database and their wavelet subbands. Figure 6.3a- 6.3c correspond to the original image whereas Figure 6.3d-6.3f and Figure 6.3g- 6.3i correspond to the two enhanced versions. Figure 6.3d has been enhanced by Top Hat Transformation based method [200] whereas Figure 6.3g has been enhanced by Global Histogram Equalization (GHE) method [201] showing a lot of over-enhancement. In order to measure changes in structure, we have evaluated SMO for each image and their subbands and to find variations in the lightness order, we have computed LOM for all of them. From the Figure 6.3, we can observe that LOM and SMO fail to distinguish correctly when applied to the two enhanced images i.e. they both give better results for Figure 6.3g as compared to Figure 6.3d in contrast to subjective score, MOS. However, SMO applied to lowpass subband of the images and LOM applied to bandpass subband both give the correct differentiation.

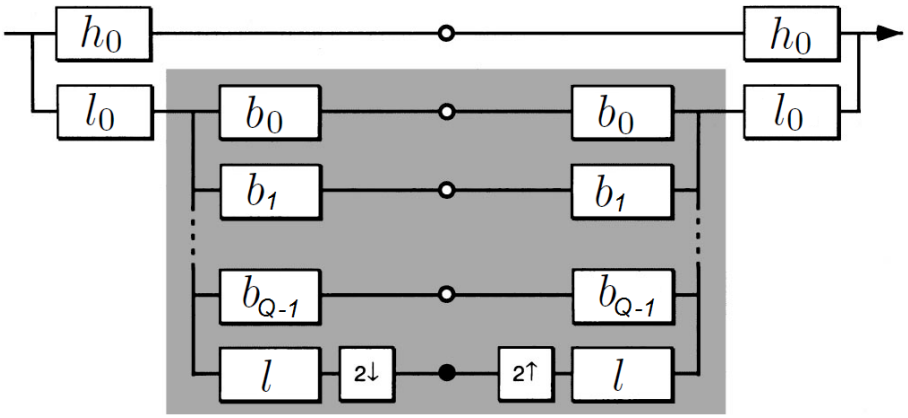
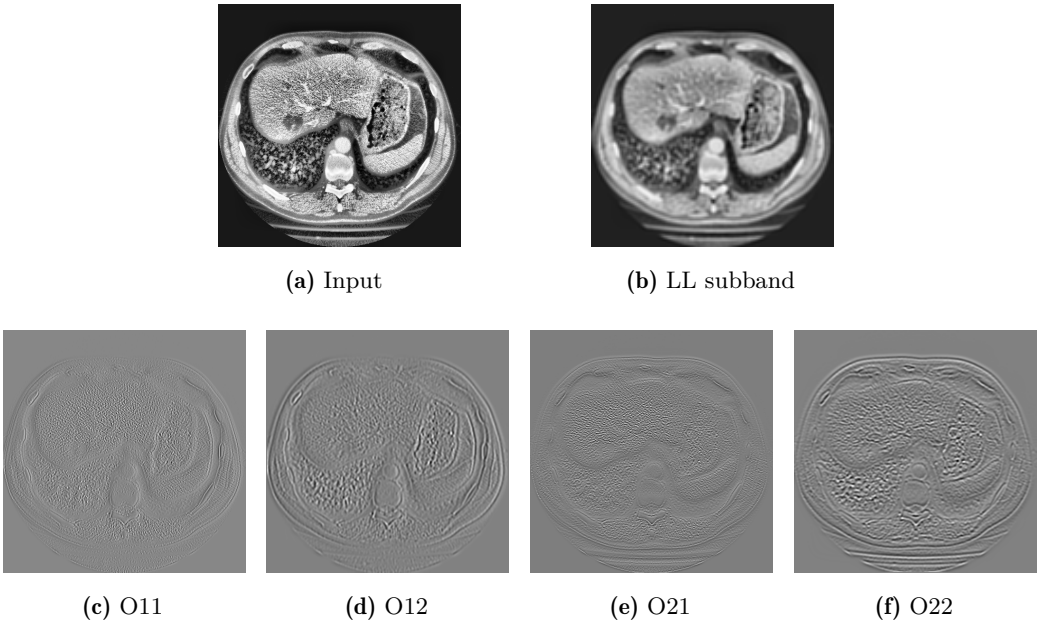


Figure 6.4: Steerable pyramid decomposition [202]

For this work, we have considered steerable pyramid wavelet transform [203] for image decomposition before performing CEE. The Steerable Pyramid is a linear multi-scale and multi-orientation image decomposition, which first splits the image into a high and low frequency parts and then applies sequentially bandpass filtering to the low frequency image part. Figure 6.4 shows the process of steerable pyramid transformation. It can capture texture variations in both intensity and orientation. The use of steerable pyramid decomposition is well-motivated for quality assessment tasks, thanks to its similarity with characteristics of the Human Visual System (HVS) and its translation invariance property[203] in contrast to classical wavelet. Figure



(a) Input **(b)** LL subband

(c) O11 **(d)** O12 **(e)** O21 **(f)** O22

Figure 6.5: CT image with steerable pyramid decomposition with 2 scales and 2 orientations

6.5 shows results of applying steerable pyramid transform to a CT image at 2 scales and 2 orientations.

6.4 Proposed MCCEE based Evaluation for CT images

In this section, we first present our proposed MCCEE metric that is based on wavelet decomposition and the use of multiple criteria. This is followed by a detailed explanation of our proposed task-based evaluation, which is needed in combination with the metric for CEE in CT images.

6.4.1 Proposed MCCEE Metric

Based on the criteria highlighted in the Section 6.2, we propose a new CEE metric that combines the four metrics mentioned. However, instead of evaluating the criteria for enhancement for an image directly, we propose in this work to first decompose the image into its wavelet counterpart. This allows to exploit the image features at different scales more effectively in the computation of the global CEE. Indeed, a better analysis of the CE side effect should be achieved when separating the different subbands of the input image. For this reason, we have chosen to evaluate SMO for bandpass subband whereas LOM for the lowpass subband of the wavelet decomposed image.

For our proposed MCCEE metric, we first decompose the image using the steerable pyramid transform at 2 scales and 2 orientations. Once the image is decomposed, we make use of only

lowpass subband and only one orientation (bandpass) subband of the second scale for applying LOM and SMO respectively. LOM evaluated for lowpass subband image and SMO computed for bandpass subband image are then combined along with AMBE and EMEE for the whole image to give the following contrast enhancement criteria feature vector,

$$\mathbf{f}^{(MCCEE)} = \Big(LOM_{LL}, SMO_{HL}, AMBE, EMEE\Big) \tag{6.5}$$

It is worth noticing here that all these metrics in $\mathbf{f}^{(MCCEE)}$ work on luminance and do not cater to the color aspect explicitly. However, this is not a cause of concern since most of the CE methods consider only the luminance channel when processing the pixel values. In these methods, the transformed luminance component is combined with the original chrominance components. Hence, these CEE criteria still capture all the most important aspects of evaluation even for CE of color images.

In order to find the value for MCCEE, we then train a Support Vector Regressor (SVR) with the feature vectors $\mathbf{f}^{(MCCEE)}$ from the data of CEE images and their corresponding human subjective scores. In order to train SVR, we have used 80% of the dataset while the remaining 20% has been used for validation. Once the SVR is trained, MCCEE value for a new image can then be predicted using its $\mathbf{f}^{(MCCEE)}$ as the input to SVR. However, since we do not have human scores for CT data, we propose to do a task-based CEE where the quality of results from the subsequent task is used as a benchmark for training MCCEE. This method is described in more detail in the next section.

6.4.2 Task-Based Contrast Enhancement Evaluation

For evaluation of contrast for CT images, we propose to do a task-based evaluation due to absence of labeled data. In this approach, contrast enhancement is evaluated based on the performance of the subsequent tasks such as segmentation. It is important to remember that the main motivation for enhancing the contrast of low contrast CT images is to facilitate tumor segmentation. Hence, in this context, a better CE method would be the one which gives a better segmentation result with the same segmentation algorithm. Moreover, since there are existing medical databases containing the segmentation ground-truth, we can evaluate the segmentation quality for each of the segmentation results obtained.

In the context of medical image segmentation, deep learning based methods have attracted a great attention in the last years [204]. However, these methods need extensive amount of data to train the network. Since only a few liver tumor images were available for this study (out of

10 patients data), it was not feasible to apply deep learning method in our case. We therefore selected gradient driven Seeded Region Growing (SRG) algorithm for our task-based CEE. In the following sections, we describe different steps of task-based CEE in more detail.

6.4.2.1 Contrast Enhancement of CT images

Low contrast CT images are first enhanced using some state-of-the-art CE methods. Two different sets of CE algorithms were used in this work for two different experiments. For the first experiment, four different CE methods were used. All of these were histogram based methods where [205] and [206] employ optimization based histogram processing and [192] applies cross-modality guidance based histogram specification for contrast enhancement. The fourth method is an optimized guided CE approach. We denote these methods as Averaging Histogram Equalization (AVHEQ) [205], Histogram Equalization with Maximum Intensity Coverage (HEMIC) [206], Cross-Modality Guidance-based enhancement (CMGE) [192] and OPTimized Guidance based Contrast Enhancement (OPTGCE). Both CMGE and OPTGCE make use of corresponding MR images for guidance-based enhancement i.e. MR images are used to guide the enhancement process in CT images.

For the second experiment, four other state-of-the-art CE methods were used. These were Brightness Preserving Dynamic Histogram Equalization (BPDHE) [207], Dynamic Histogram Equalization [208] Contrast-Limited Adaptive Histogram Equalization (CLAHE) [209] and Tune Brightness Controlled single scale Retinex (TBCSSR) [210]. None of these methods employ guided CE.

6.4.2.2 Seeded Region Growing Segmentation

For the reasons explained earlier, we have opted for the region based image segmentation method, that is, the SRG algorithm [211]. Region growing algorithm maps the image pixels into sets of grouped pixels based on a criterion specified by the features of local neighborhood of the pixels. The process of SRG is depicted in Figure 6.6, where starting from seed pixel the region grows as long as the specified criteria is met. Seeded Region Growing is one of the simple yet widely accepted segmentation algorithm particularly in medical images [212, 213, 214]. Segmenting critical structures in medical images requires higher accuracy compared to other type of images. Completely automated segmentation procedures may work well on natural images, but segmentation in medical images needs significant caution since the results are later used for diagnosis and occasionally for surgical planning as well. Several studies report the use of seeded region growing for segmenting medical images compared to more sophisticated approaches

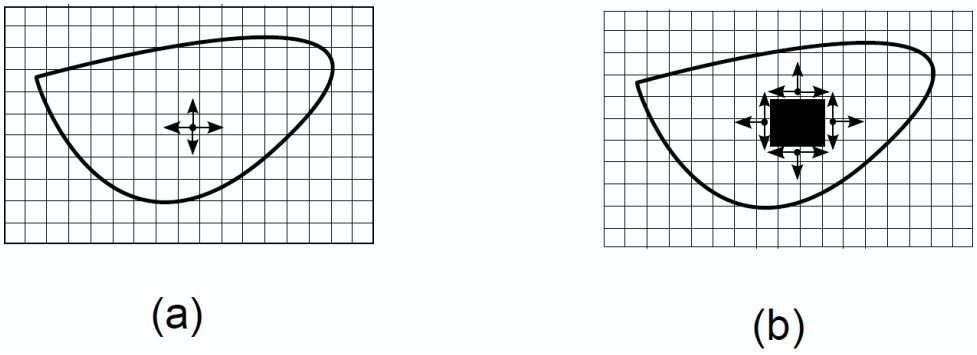


Figure 6.6: Seeded Region Growing (a) Start of Region Growing from seed pixel (b) Growing Process after a few steps

[212, 213, 214]. The governing principle of SRG inherently assumes that pixels from specific region share same values, which often does not work well on all types of images. We use gradient driven seeded region growing in this work to address this limitation. The criteria for region growth is specified by a cost function that extracts the features of pixels surrounding the seed that is gradient magnitude and direction when the region evolves. A constraint that may limit the utility of gradient driven SRG is the execution time; therefore, we use its parallel implementation as proposed in [215].

The basic objective of this research work is not to propose a novel and efficient segmentation scheme, rather it is to present a contrast enhancement evaluation method for CT images based on segmentation results. We therefore selected a simple parallel SRG scheme, which is frequently applied in segmenting critical structures in medical images. The segmentation evaluation method is discussed in the next section.

6.4.2.3 Segmentation Evaluation

We used three assessment metrics to quantitatively evaluate and validate the performance of the segmentations. These include Dice, Positive Predictive Value (PPV) and Hausdorff distance (with euclidean distance). Several metrics have been proposed in the past to evaluate the performance of segmentation algorithms such as intensity based, shape based and distance based. One of the challenges in medical image segmentation assessment is that the object of interest constitutes smaller part of image, and therefore the assessment methods are biased to yield more weightage to specificity compared to sensitivity. Distance based metrics such as Hausdorff distance are capable of detecting data outliers and better handle the inconsistent data in two datasets whose Hausdorff distance is being investigated. Intensity based assessment approaches may fail in such scenarios.

Dice Coefficient (DICE)

The Dice coefficient (DICE) [216] is the most commonly used metric for validation of medical image segmentation. It is used to find the overlap between the ground-truth segmentation S_g and the test segmentation S_t using

$$DICE = \frac{2|S_g \cap S_t|}{|S_g| + |S_t|} \tag{6.6}$$

where $|S_g|$ and the $|S_t|$ are the cardinalities of the two sets.

Positive Predictive Value (PPV)

Positive Predictive Value (PPV) is simply defined as the ratio between True Positives (TP) and the sum of TP and False Positives (FP)

$$PPV = \frac{\sum TP}{\sum TP + \sum FP} \tag{6.7}$$

Hausdorff Distance (HD)

The Hausdorff Distance (HD) is a spatial-distance based metric used to evaluate dissimilarity between two segmentations. Like other distance-based measures, the spatial distance is measured using spatial positions of the voxels. For two finite point sets, HD is defined in terms of directed Hausdorff distance $h(A, B)$ as

$$HD(A, B) = \max(h(A, B), h(B, A)) \tag{6.8}$$

where

$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{6.9}$$

with $||.||$ is some norm like the Euclidean distance. A smaller value of HD implies better segmentation results.

6.5 Experimental Results and Discussion

Before testing our proposed MCCEE metric for the task-based CEE, we first perform a comparative analysis with other state-of-the-art CEE metrics using human scores from labeled natural CEE image databases. Thereafter, we perform two experiments for task-based CEE of CT images. In the first experiment, we validate the correlation between segmentation improvement and contrast enhancement. For this experiment, we have chosen optimization based methods

for contrast enhancement. For evaluation of the segmentation results, we use the three metrics described in Section 6.4.2.3 in addition to doing a qualitative analysis. Finally, we also evaluate MCCEE and two other CEE metrics to highlight the effectiveness of our metric.

In the second experiment, we use a different CT database and four different CE methods. Here, we do a more detailed comparative analysis similar to the one done for labeled natural CEE databases in the first step. This helps to validate whether the results from the subjective evaluations could also be extended to task-based CEE.

6.5.1 Results with Existing Natural Image Databases

As a first step, we evaluated the performance of our new metric on two different CEE databases namely CEED2016 [78, 196] and CCEID [197]. This is because both of these databases contain subjective scores from human observer experiments.

CEED2016 is a database dedicated to contrast enhancement evaluation. It is composed of 30 reference images, each of which is enhanced by 6 representative CE methods resulting in a total of 180 enhanced images. The employed CE algorithms used are Adaptive Edge Based Contrast Enhancement (AEBCE) [80], Contrast Limited Adaptive Histogram Equalization (CLAHE) [209], Discrete Cosine Transform based (DCT) [217], Global Histogram Equalization (GHE) [201], Top Hat Transformation based (TOPHAT) [200], and Multi-scale Retinex (MRETINEX) [218].

CCEID database on the other hand contains 26 pictorial reference images. To each of these reference images, four contrast enhancement methods have been applied. These methods are CLAHE [209], s-shaped contrast correction [219], Retinex and Natural Rendering of Colour Image using Retinex (NRCIR) [220].

In order to evaluate the performance of our MCCEE metric, we have calculated three different correlation coefficients to check correlation between our metric and the subjective scores. These coefficients are Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC) and Kendall Rank-Order Correlation Coefficient (KROCC). Before evaluating the correlation coefficients, the scores from metrics are passed through a 5-parameter logistic function. Moreover, in order to check the effectiveness of our metric, we have compared the coefficient values for our metric to those obtained with other CEE metrics. In total, we have compared the results of the proposed metric with 14 other selected CEE metrics from [78]. These are AMBE, Absolute Measure of Enhancement (AME), Absolute Measure of Enhancement by Entropy (AMEE), EME, EMEE, Discrete Entropy (DE), EC, IEM, LOE, LOM, Root Mean Square Contrast (RMSC), RSE, Second Derivative like MEasurement (SDME) and

Table 6.1: Results with CEED2016 dataset

CEE metric	PLCC	SROCC	KROCC
AMBE	0.229	0.254	0.179
AME	0.272	0.216	0.145
AMEE	0.349	0.329	0.221
DE	0.186	0.073	0.048
EC	0.310	0.236	0.159
EME	0.428	0.302	0.208
EMEE	0.505	0.399	0.273
IEM	0.573	0.384	0.267
LOE	0.377	0.387	0.262
LOM	0.713	0.690	0.498
RMSC	0.436	0.376	0.254
RSE	0.201	0.166	0.115
SDME	0.283	0.241	0.167
SMO	0.567	0.511	0.351
Proposed	**0.855**	**0.830**	**0.648**

SMO. For further details on these metrics the readers can refer to the work by Qureshi et al. [78].

Table 6.1 and Table 6.2 show the correlation results of all the metrics for CEED and CCEID databases, respectively. The results highlighted in bold correspond to the best values for each correlation coefficent for the database. It is clear from these results that the proposed metric outperforms all others significantly as a CEE metric. For CEED database, we see a significant increase of 0.14 for PLCC and SROCC, and an increase of 0.15 for KROCC from the second best performing metric of LOM. For CCEID database, it is worth noting that the correlation values are lower for all the metrics. As suggested by the authors of this database [197], these lower values are due to the already good contrast of the reference images. However, even for CCEID database, we see a significant improvement of at least 0.11 for all correlation coefficients from the second best performing metric (LOE for PLCC and KROCC and AMBE for SROCC).

Table 6.2: Results with CCEID dataset

CEE metric	PLCC	SROCC	KROCC
AMBE	0.428	0.418	0.290
AME	0.247	0.122	0.085
AMEE	0.257	0.220	0.157
DE	0.113	0.124	0.091
EC	0.091	0.180	0.132
EME	0.331	0.155	0.114
EMEE	0.346	0.295	0.203
IEM	0.328	0.208	0.138
LOE	0.474	0.412	0.282
LOM	0.370	0.267	0.198
RMSC	0.465	0.371	0.258
RSE	0.08	0.024	0.019
SDME	0.107	0.077	0.062
SMO	0.275	0.272	0.188
Proposed	**0.586**	**0.526**	**0.405**

6.5.2 Effects of CE on Segmentation Performance for Task-Based CEE

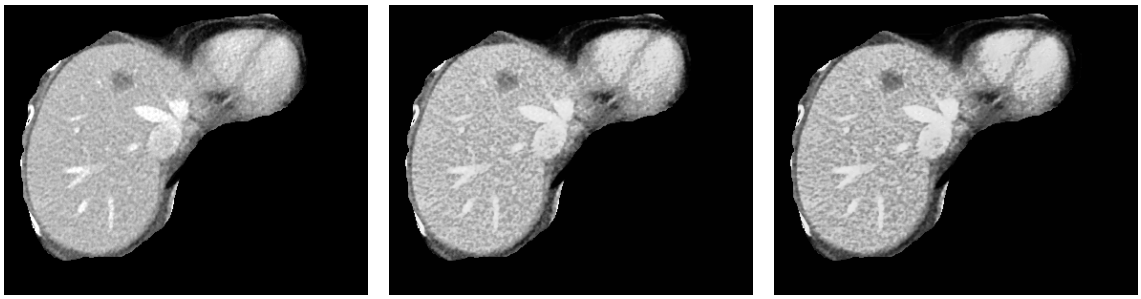
In this section, we explore the effects of CE on subsequent segmentation performance. First, we explain the dataset used in the experiment followed by the results obtained using different methods[192, 206, 205]. Finally, we do a qualitative and quantitative analysis of the results to validate the effects of CE on segmentation.

6.5.2.1 Contrast Enhancement

The data used in the proposed research work is provided by Intervention Center, Oslo University Hospital. In this dataset, we have Liver CT and MR images of the same patient. This is a necessary requirement for the two guided CE methods. We have not done image to image registration as it is not required for global enhancement methods. We performed CE using the four methods described in Section 6.4.2.1 on 10 patients data constituting of 99 CT-MR image pairs (containing tumors). The images from different volumes are of different spatial size (such as 512×512, 360×240) within the grayscale range [0, 255]. In medical image processing tasks

such as registration, segmentation and enhancement, the processing is often restricted to the particular organ and the nearby organs are removed from the medical images [221, 222]. The liver CT image for instance contains nearby organs such as heart, and we are not interested in enhancing the contrast of heart for this work, therefore the liver area is cropped and proposed method is applied only on liver region in CT images.

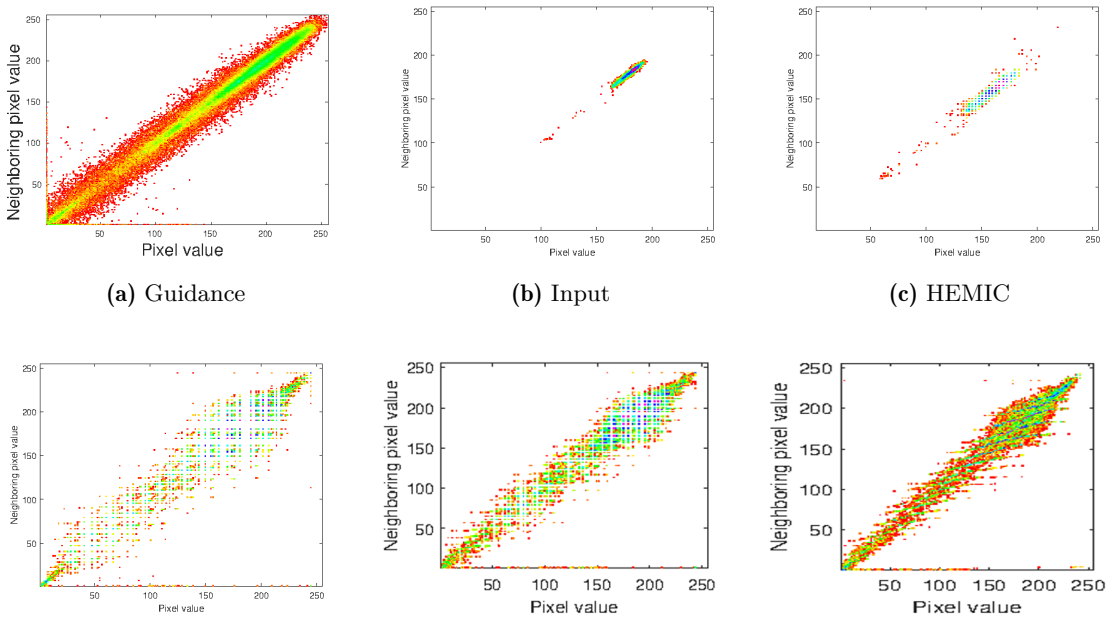
(a) Guidance **(b)** Input **(c)** HEMIC



(d) AVHEQ **(e)** CMGE **(f)** OPTGCE

Figure 6.7: Enhancement Results from state of the art methods

The input image in Figure 6.7b has low contrast as can be perceived visually. The image enhanced using HEMIC does not show noticeable contrast improvement even in terms of visual perception. Although applying enhancement proposed by CMGE expands the dynamic range of the image, yet this dynamic range is still far from ideal considering the range of guidance image. Contrary to that, the methods AVHEQ and HEMIC stretch the dynamic range of enhanced images over nearly the entire dynamic range. However, considering the co occurrence matrix plot of AVHEQ in Figure 6.8d, it can be observed that although the width of ellipse expands along the diagonal; the plot shows significant gaps among the pixel pairs and consequently compactness of the plot is lost. The plot of the image enhanced using OPTGCE in Figure 6.8f reflects uniform and compact distribution of the pixel pairs. Furthermore, it approximates plot of the guidance image in evenly distributing the pixel pairs. Hence, from empirical analysis, OPTGCE is the best in terms of improving contrast of input CT image.



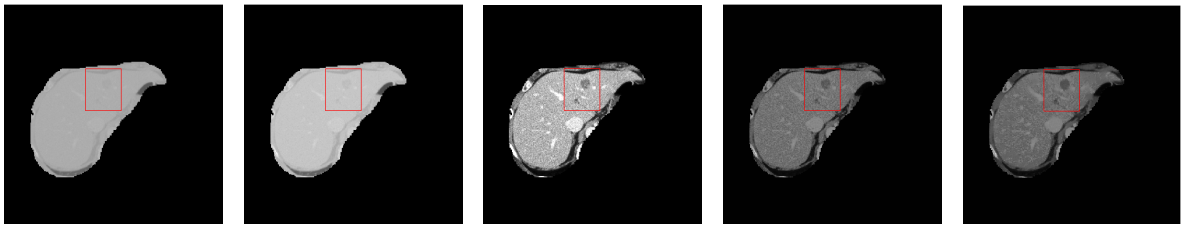
(a) Guidance **(b)** Input **(c)** HEMIC

(d) AVHEQ **(e)** CMGE **(f)** OPTGCE

Figure 6.8: Corresponding GLCM plots of guidance, input and enhanced images (red to blue decreasing pixel pair values)

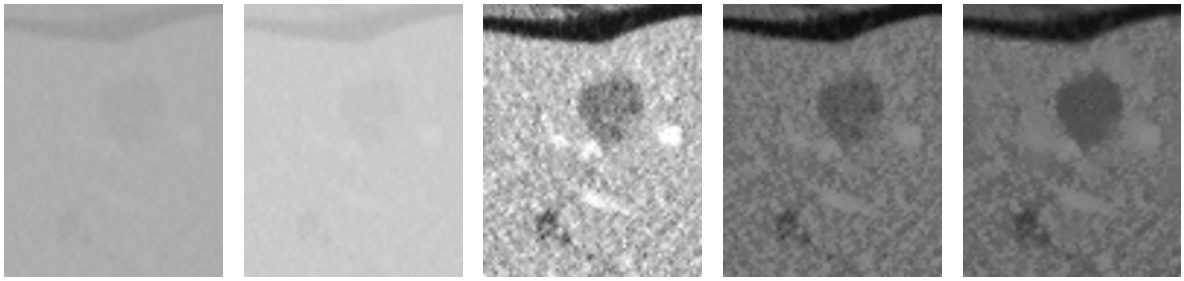
6.5.2.2 Tumor Segmentation

The ground truth for this experiment is also provided by Intervention Center, Oslo University Hospital. The segmentation results are demonstrated without applying any kind of post-processing such as morphological region filling or closing; better segmentation results could be obtained if appropriate post-processing was applied on the segmented images. Since the objective here is to enhance images to facilitate the subsequent segmentation phase, we are interested in retaining the boundary of tumor while segmenting them. The tumor can hardly be visualized in the input images in Figure 6.9a without applying a suitable quality enhancement process. Figure 6.9 shows the enhanced images that we provide as input to SRG algorithm whereas Figure 6.11 shows the corresponding segmentation results. Images with tumors in Figure 6.9e enhanced using OPTGCE method are clearly visible with tumor edges better discriminated from liver parenchyma. It can be visually inferred that the segmentation on images enhanced using proposed method is closer to ground truth in terms of structural proximity. The reason for superior performance of SRG on OPTGCE method could be observed in the enlarged view shown in Figure 6.10e. We can observe that the tumor area is more homogeneous (similar intensity values of the neighboring pixels), whereas areas composed of non-homogeneous intensities can be observed for other methods such as those illustrated in Figures 6.10d and 6.10c. Overall, segmentation does not work well on



(a) Input **(b)** HEMIC **(c)** AVHEQ **(d)** CMGE **(e)** OPTGCE

Figure 6.9: Enhancement results of different state of the art methods



(a) Input **(b)** HEMIC **(c)** AVHEQ **(d)** CMGE **(e)** OPTGCE

Figure 6.10: Tumor area enlarged from enhanced images

images enhanced using the other methods. Consider 6.11d for instance; significant part of the tumor is not included in the resulting segmentation. Another point worth mentioning here is the application of SRG on input image without any kind of enhancement. Very small part of tumor area is segmented as shown in Figure 6.11a for instance. To verify this qualitative analysis, we also perform the quantitative analysis next.

Gradient driven SRG algorithm is applied on three volumes containing a total of thirty tumors. Among the numeric results in Table 6.3, PPV values in general are greater than 0.8 for all the segmentations. Since PPV computes ratio between number of pixels correctly classified as tumors to the sum of number of pixels correctly classified and the non-tumor pixels wrongly classified

(a) Input **(b)** HEMIC **(c)** AVHEQ **(d)** CMGE **(e)** OPTGCE **(f)** Ground-truth

Figure 6.11: Tumor segmentation applied on enhanced images

as tumors, this value gives similar values to all the methods. It can be observed from Figures 6.11 that all the segmentations applied on the images do not include many non-tumor pixels in the resultant segmentations when compared against ground truth. Although Dice yields better scores for segmentation applied on images enhanced using OPTGCE method, the overall Dice scores are low. Dice similarity metric gives higher value to the terms that compute intersection between true positives in segmentation under test and ground truth. For OPTGCE approach, although the segmented area lies within the Ground-truth (GT) segmentation yet it does not completely overlap with it, hence leading to the lower Dice scores. We did not apply any kind of post-processing to segmentation results, therefore, the resulting segmentations contain rough edges, discontinuities and nonuniformity in segmented tumors. It is worth mentioning here that Dice score and Hausdorff distance consistently show lower scores when the SRG algorithm is applied on the input images without any kind of enhancement. The average values of Hausdorff distance, Dice score and PPV obtained for each volume are shown in Table 6.3. The segmentation in case of OPTGCE method consistently achieves lower Hausdorff distance values for all the three volumes. CMGE is the second best while HEMIC ranks lowest in all the three cases tested.

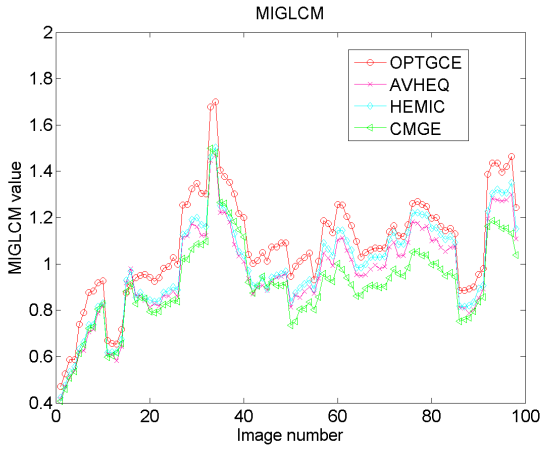
Table 6.3: Comparison of different segmentation assessment method for enhancement Results

Methods	Input w/o processing			HEMIC [206]			AVHEQ [205]			CMGE [192]			OPTGCE		
	PPV	**Dice**	**Haus.**	**PPV**	**Dice**	**Haus.**	**PPV**	**Dice**	**Haus.**	**PPV**	**Dice**	**Haus.**	**PPV**	**Dice**	**Haus.**
vol1	**0.67**	0.16	14.21	0.63	0.26	23.50	0.66	0.17	17.00	0.56	0.36	14.60	**0.67**	**0.46**	**9.92**
vol2	**0.90**	0.14	8.66	0.83	0.26	8.26	0.86	0.27	7.70	0.88	0.38	7.30	**0.90**	**0.45**	**5.90**
vol3	0.85	0.31	16.18	0.85	0.36	15.78	0.86	0.42	14.00	0.83	0.48	12.47	**0.91**	**0.57**	**10.50**

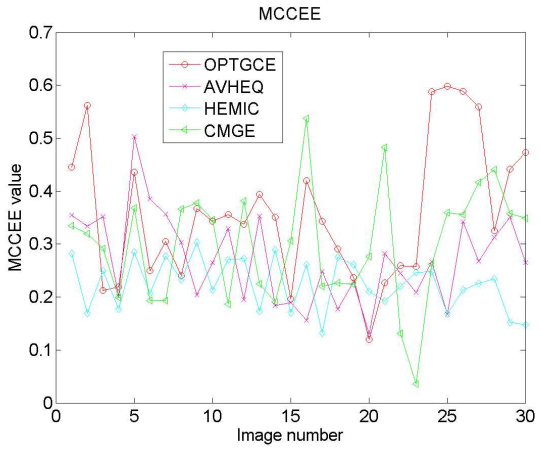
6.5.2.3 CEE Results

The motivation of a good CE method such as OPTGCE is to emphasize appearance of specific structures in the image and convey maximum structural information so as to facilitate the tumor segmentation phase. To this end, we have chosen two different CEE metrics in addition to our proposed MCCEE to evaluate and compare the quality. The first metric is mutual information based no reference metric called MIGLCM [223]. This metric offers quantitative criteria that examines the changes in statistical features joint entropy and mutual information acquired from the co-occurrence matrix of the original and enhanced images.

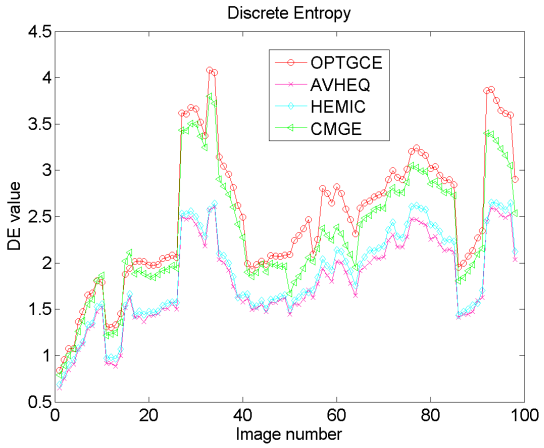
Besides MIGLCM, we have used Discrete Entropy (DE) measure which is often used in QA of medical image enhancement [205]. For MCCEE, since we have no subjective scores for CE methods, we have selected Dice score for task-based training and validation. This choice is arbitrary and any of three metrics from PPV, Dice and HD could have been selected as they



(a) Assessment using MIGLCM



(b) Assessment using MCCEE



(c) Assessment using DE

Figure 6.12: Quantitative assessment of different enhancement methods

all perform similarly well as shown in the previous section. It should be noted that MCCEE here is applied on data of three patients only since it needs a training phase and segmentation is applied on 3 volumes.

The results for the three metrics on all images are shown in Figure 6.12 whereas Table 6.4 lists the median values for these metrics from all images. From the figure and the table, we can observe that OPTGCE demonstrates the best performance. For MCCEE and DE, CMGE, HEMIC and AVHEQ are ranked low overall by the two QA metrics. In case of MIGLCM, HEMIC is ranked as the second best and CMGE gives the poorest results. Overall, we observe that OPTGCE shows the best performance from all 3 quantitative methods chosen. However, when it comes to ranking all the methods correctly, we see that only MCCEE gives the precise ranking by giving the lowest score to HEMIC.

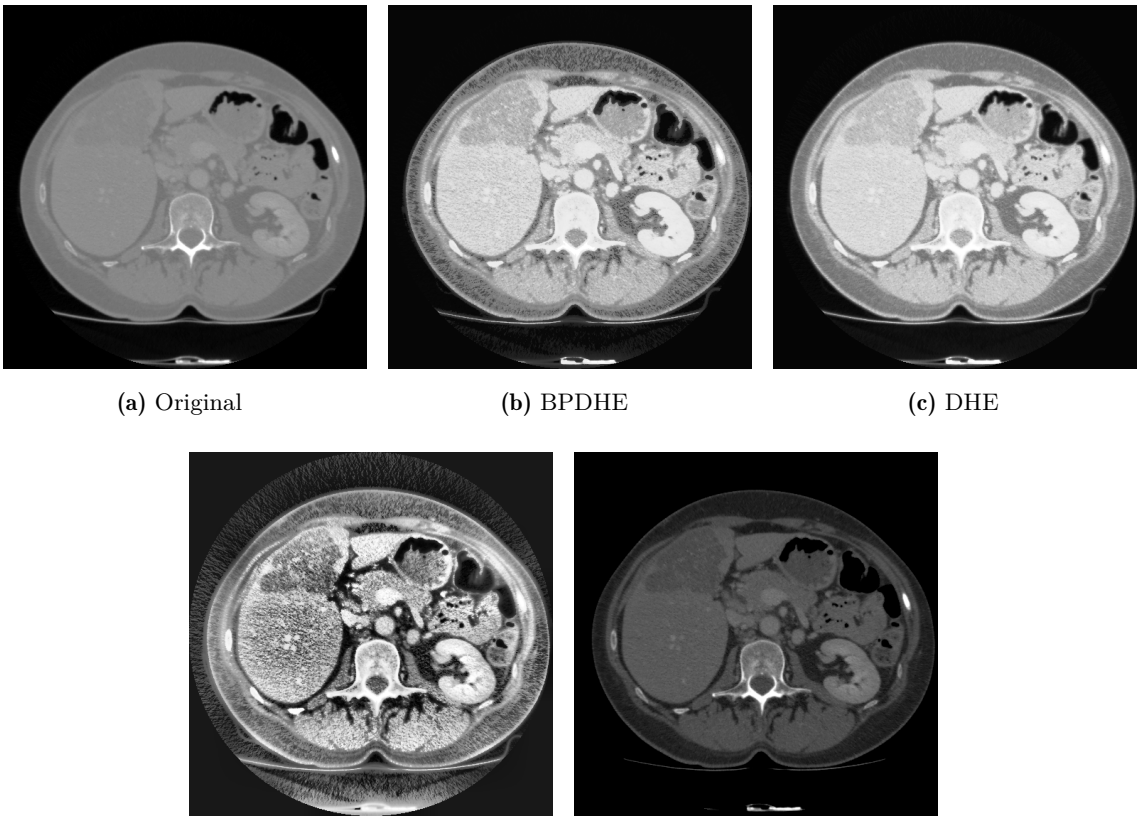
To sum up, with this experiment, we found out qualitatively and quantitatively that a better CE method does improve the performance of the subsequent task of tumor segmentation. This was verified even by existing CEE metrics which gave the best scores for the best method of OPTGCE. However, we also observed that the existing CEE metrics also fail sometimes to rank the CE methods correctly as they gave better scores to HEMIC in comparison to AVHEQ although the inverse is correct as can be seen from Figure 6.10. Finally, it was also observed that segmentation evaluation methods were consistent with each other and hence we can choose any of these metrics to train our MCCEE metric. For this work, we have selected the most commonly used Dice score for training our proposed MCCEE metric. In the next section, we perform a detailed comparative analysis with other CEE metrics using a different dataset and a different set of CE methods. Using the conclusions derived in this section, we also evaluate the correlation coefficients using Dice instead of subjective scores as the true labels.

Table 6.4: Median CEE metric values for different methods

CEE metric	**AVHEQ**	**HEMIC**	**CMGE**	**OPTGCE**
MIGLCM	0.956	0.984	0.908	**1.070**
MCCEE	0.266	0.229	0.313	**0.343**
DE	1.680	1.730	2.160	**2.470**

6.5.3 Results with CT images using Task-Based CEE

In this section, we provide a comparative analysis of MCCEE with other state-of-the-art methods. However, since there is no existing database available publicly for medical data with subjective scores, we have used a task-based approach where contrast enhancement is evaluated based on the performance of the subsequent tasks such as segmentation. Hence, we first evaluate a contrast enhancement method by judging the quality of tumor segmentation from the enhanced image. Thereafter, these results are correlated to the CEE metric scores. For this experiment, we have selected 7 liver CT images from the commonly used public 3DIRCADb database [224] and have then applied 4 different contrast enhancement methods on each of them. These methods are Brightness Preserving Dynamic Histogram Equalization (BPDHE) [207], Dynamic Histogram Equalization [208] Contrast-Limited Adaptive Histogram Equalization (CLAHE) [209] and Tune Brightness Controlled single scale Retinex (TBCSSR) [210].



(a) Original **(b)** BPDHE **(c)** DHE

(d) CLAHE **(e)** TBCSSR

Figure 6.13: Enhanced Liver CT images

Figure 6.13 shows an example liver CT image with tumor and its four enhanced versions. For segmentation of tumors, we have used the simple region growing technique [211]. In order to

judge the quality of tumor segmentation, we have used Dice score [216] which is used to find the overlap between the ground truth segmentation and the output of segmentation algorithm. For MCCEE training on this dataset, we also used Dice score as the ground truth instead of the subjective scores.

Table 6.5: Results with Liver CT dataset

CEE metric	PLCC	SROCC	KROCC
AMBE	0.511	0.525	0.407
AME	0.500	0.485	0.354
AMEE	0.137	0.025	0.032
DE	0.358	0.366	0.249
EC	0.448	0.438	0.291
EME	0.580	0.580	0.402
EMEE	0.626	0.605	0.397
IEM	0.350	0.368	0.270
LOE	0.334	0.342	0.243
LOM	0.182	0.194	0.153
RMSC	0.448	0.433	0.280
RSE	0.298	0.250	0.196
SDME	0.003	0.178	0.111
SMO	0.319	0.271	0.175
Proposed	**0.832**	**0.700**	**0.600**

For the medical dataset, we have similarly evaluated PLCC, SROCC and KROCC for all metrics to compare the ranking performance. Since there are no subjective scores to use as reference, we have made use of the Dice scores as the reference scores, which give a good indication of the preprocessing task of CE and its effect on the segmentation results as shown in the last section. Table 6.5 shows the results for the liver CT dataset. Here, again we observe a substantial improvement in assessment of CE with the proposed MCCEE measure as compared to the other state-of-the-art metrics. For PLCC and KROCC, we observe a big improvement of around 0.2 in correlation coefficient values as compared to the second best-performing metrics (EMEE for PLCC and AMBE for KROCC) whereas for SROCC this improvement is around 0.1 as compared

to second-best EMEE.

6.6 Conclusion and Perspectives

Through this study, we have shown that by collaboratively combining different CEE metrics we can capture the major effects that condition the perceptual quality of the processed image. The major contribution of this work is to show that using a simple SVR based learning scheme with the representative CE criteria allows to achieve relatively large correlations with the subjective assessment on reference databases. Moreover, we also proposed a novel task-based assessment strategy for CEE where instead of using subjective scores, we made use of the results of the subsequent task of segmentation. It is clear that by adopting an approach that would be in line with the current deep learning trend, we could produce more consistency with the subjective assessment of the reference. But, this would require large databases dedicated to this particular case of image quality assessment, which is very different from what is generally known in the image quality assessment community. This is one of the lines of work in perspective both experimentally and theoretically in the search for an architecture based on deep learning that is explainable and coherent with a scientific approach justified for each stage of the statistical learning system.

CHAPTER 7

Conclusions and future work

In this chapter, we draw some conclusions from the work that has been done in this thesis and discuss some perspectives for future work. First, in Section 7.1 we provide an overall summary of the work achieved in the context of our original thesis objectives and the conclusions which can be derived from them. Then, in Section 7.2, we provide some ideas for improvement and extension of this work.

7.1 Conclusions

The emergence of new medical imaging modalities have brought about a revolution in diagnostic medicine. It has also encouraged an increase in the use of imaging for performance of surgery. For this reason, we have seen a boost in the number of surgeries being performed using laparoscopy and image-guidance in the recent times. However, applications as critical as diagnostics and surgery, bring forward their own sets of challenges. One of these challenges is the preservation of a good quality of images throughout. In order to achieve this, there is a need of a continuous assessment of the quality of images. Furthermore, processes like pre-operative surgical planning and intra-operative guidance rely on processing tasks like image enhancement, image registration and image segmentation. For the success of surgery, it is equally important to monitor the performance of these tasks. The most feasible way to assess the quality of both the acquired images and the processing tasks is to do it objectively in an automated manner. However, existing research which tackles these problems in the medical context is very limited. This thesis has taken on these challenges and has provided contributions to the quality assessment of medical

imaging modalities and processing tasks.

In the following paragraphs, we provide a brief summary of the work done in this thesis and their conclusions, before listing the future perspectives in the next section.

First of all, we have investigated the use of joint statistics for quality assessment of stereoscopic images. To this end, we have proposed to exploit inter-view and intra-view dependencies in the wavelet representations of the stereo pairs by proposing bivariate and multivariate generalized Gaussian models for texture feature extraction. We have tested our method on an existing stereoscopic quality database and have shown that these models provide slight performance improvement especially for some specific distortions like JPEG, white noise and fast fading.

For the next contribution, we have constructed a new quality database of 2D laparoscopic videos with subjective scores called the LVQ database. In order to achieve this, we have first simulated distortions in some laparoscopic videos that have been selected from some public datasets. Then, subjective evaluations by expert medical observers and non-expert observers were performed on the database. Based on these observations, we have evaluated MOS for both the groups. We have also done an initial statistical analysis of the scores which has shown that there are some differences in the way experts and non-experts perceive image quality. There were larger variations found in the opinions of experts as compared to non-experts, providing an insight into the difficulty of modeling expert opinion. Furthermore, we have also seen from this work that none of the existing objective quality metrics correlated well with the subjective scores. These correlations were found to be worse for the expert group.

As the third contribution, we have proposed a deep learning based method for no-reference objective quality assessment of 2D laparoscopic videos. The network proposed in this work uses a residual network and aims at solving the dual task of distortion classification and video quality prediction. We have also done a comparison of this method with existing state-of-the-art methods and it shows that our method has the best performance. In addition to that, we have also seen that using a transfer learning approach from image distortion classification and ranking task to video quality prediction task gives a significant improvement in the correlation of the predicted scores with human scores. Moreover, our proposed temporal pooling method using a fully-connected network has also shown much better results as compared to conventional average temporal pooling.

Finally, for evaluation of contrast enhancement of CT images, we have proposed a novel assessment metric based on the use of some selected criteria. The latter are essential for any effective contrast enhancement method. We have selected some metrics to measure these criteria

and have shown that overenhancement measures among them give better results when used with wavelet decomposition. Using these criteria and a wavelet decomposition, we have then applied a machine learning based approach for evaluation of the final metric. Our results on different databases show that our metric significantly outperforms all other existing metrics. Furthermore, we have also concluded that a task-based contrast enhancement strategy for CT images, that uses the performance results of subsequent processing tasks like segmentation, also gives excellent results. This strategy can be used when there are no subjective scores for the contrast-enhanced images are available.

7.2 Future work

The research work related to quality assessment of multiple modalities used in medical imaging, carried out in this thesis can be further improved and/or extended in several directions. Among them, we can mention the following ones.

- Extend the proposed LVQ database by including more different distortions like specular reflection and foggy lens or haze. In this regards, a very important aspect is also to consider videos having multiple simultaneous distortions. Another vital aspect would be to add videos with real distortions rather than synthetic ones in the database.
- Develop large quality databases for other medical modalities like CT, MRI and Ultrasound to inspire development of deep learning based methods for objective assessment. This must also include specialized databases for the contrast enhancement evaluation.
- Extend and adapt the proposed statistical based stereo image quality assessment to the context of video quality assessment.
- Evaluate the effectiveness of fusion-based methods for combining multiple criteria identified for CEE in Chapter 6. Furthermore, other kinds of image decomposition may also be exploited rather than the retained wavelet decomposition.
- Extend the method based on multiple criteria for CEE to evaluation of segmentation. This can be done using a similar approach by identifying important segmentation criteria and then using machine learning method to predict the quality of segmentation.
- Implement the complete quality monitoring pipeline of Figure 4.1 using the proposed quality assessment and CEE methods and by incorporating suitable enhancement methods for removing distortions.

Bibliography

[1] F. Patrona, I. Mademlis, F. Kalaganis, I. Pitas, and K. Lyroudia, "Stereoscopic medical data video quality issues," *Journal of Medical Imaging*, vol. 3, no. 2, pp. 025 501–025 501, 2016.

[2] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.

[3] K.-H. Thung and P. Raveendran, "A survey of image quality measures," in *2009 international conference for technical postgraduates (TECHPOS)*. IEEE, 2009, pp. 1–4.

[4] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," *Proceedings of the National Academy of Sciences*, vol. 90, no. 21, pp. 9758–9765, 1993.

[5] X. He and S. Park, "Model observers in medical imaging research," *Theranostics*, vol. 3, no. 10, p. 774, 2013.

[6] L. Zhang, C. Cavaro-Ménard, and P. Le Callet, "An overview of model observers," *Innovation and Research in BioMedical engineering (IRBM)*, vol. 35, no. 4, pp. 214–224, 2014.

[7] L. Zhang, C. Cavaro-Ménard, P. Le Callet, and J.-Y. Tanguy, "A perceptually relevant channelized joint observer (pcjo) for the detection-localization of parametric signals," *IEEE transactions on medical imaging*, vol. 31, no. 10, pp. 1875–1888, 2012.

[8] M. P. Eckstein and C. K. Abbey, "Model observers for signal-known-statistically tasks (sks)," in *Medical Imaging 2001: Image Perception and Performance*, vol. 4324. International Society for Optics and Photonics, 2001, pp. 91–102.

[9] A. G. George and K. Prabavathy, "A survey on different approaches used in image quality assessment," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 14, no. 2, p. 78, 2014.

[10] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *Majlesi Journal of Electrical Engineering*, vol. 9, no. 1, pp. 55–83, 2015.

[11] B. Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 500–13, 2012.

[12] I. BT, "500-14. bt. 500: Methodologies for the subjective assessment of the quality of television images," *International Telecommunications Union: Geneva, Switzerland*, 2019.

[13] B. Series, "Methodology for the subjective assessment of video quality in multimedia applications," *Recommendation ITU-R BT*, p. 1788, 2007.

[14] R. ITU-T, "P910," *Subjective video quality assessment methods for multimedia applications*, 2008.

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

[16] X. Wang, M. Yu, Y. Yang, and G. Jiang, "Research on subjective stereoscopic image quality assessment," in *Proc. SPIE*, vol. 7255, no. 725509, 2009, p. 2009.

[17] T. C. Brown and G. L. Peterson, "An enquiry into the method of paired comparison: reliability, scaling, and thurstone's law of comparative judgment," *Gen Tech. Rep. RMRS-GTR-216WWW. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 98 p.*, vol. 216, 2009.

[18] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[19] L. L. Thurstone, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[20] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[21] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomedical Signal Processing and Control*, vol. 27, pp. 145–154, 2016.

[22] N. Sinha and A. Ramakrishnan, "Quality assessment in magnetic resonance images," *Critical Reviews™ in Biomedical Engineering*, vol. 38, no. 2, 2010.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402.

[25] J.-F. Pambrun and R. Noumeir, "Limitations of the ssim quality metric in the context of diagnostic imaging," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2960–2963.

[26] 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.

[27] M. Razaak, M. G. 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.

[28] R. Kumar and M. Rattan, "Analysis of various quality metrics for medical image processing," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 11, 2012.

[29] H. Rajagopal, L. S. Chow, and R. Paramesran, "Subjective versus objective assessment for magnetic resonance images," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 12, pp. 2422–2427, 2015.

[30] M. Fuderer, "The information content of mr images," *IEEE Transactions on Medical Imaging*, vol. 7, no. 4, pp. 368–380, 1988.

[31] L. Moraru, S. S. Moldovanu, and C. D. Obreja, "A survey over image quality analysis techniques for brain mr images," *International Journal of Radiology*, vol. 2, no. 1, pp. 24–28, 2015.

[32] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal processing letters*, vol. 17, no. 5, pp. 513–516, 2010.

[33] ——, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[34] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[35] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[36] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.

[37] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1098–1105.

[38] D. Varga, D. Saupe, and T. Szirányi, "Deeprn: A content preserving deep architecture for blind image quality assessment," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[39] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[40] 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, 2017.

[41] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[42] A. Chetouani, "A blind image quality metric using a selection of relevant patches based on convolutional neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1452–1456.

[43] M. Osadebey, M. Pedersen, D. Arnold, and K. Wendel-Mitoraj, "Bayesian framework inspired no-reference region-of-interest quality measure for brain mri images," *Journal of Medical Imaging*, vol. 4, no. 2, pp. 025 504–025 504, 2017.

[44] A. Mason, J. Rioux, S. E. 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, pp. 1064–1072, 2019.

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

[46] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE transactions on image processing*, vol. 9, no. 4, pp. 636–650, 2000.

[47] T. Küstner, P. Wolf, M. Schwartz, A. Liebgott, F. Schick, S. Gatidis, and B. Yang, "An easy-to-use image labeling platform for automatic magnetic resonance image quality assessment," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 754–757.

[48] L. Lévêque, H. Bosmans, L. Cockmartin, and H. Liu, "State of the art: Eye-tracking studies in medical imaging," *Ieee Access*, vol. 6, pp. 37 023–37 034, 2018.

[49] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 2110–2114.

[50] Y.-H. Lin and J.-L. Wu, "Quality assessment of stereoscopic 3D image compression by binocular integration behaviors," *IEEE transactions on Image Processing*, vol. 23, no. 4, pp. 1527–1542, 2014.

[51] F. Shao, K. Li, W. Lin, G. Jiang, and M. Yu, "Using binocular feature combination for blind quality assessment of stereoscopic images," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1548–1551, 2015.

[52] B. Appina, S. Khan, and S. S. Channappayya, "No-reference stereoscopic image quality assessment using natural scene statistics," *Signal Processing: Image Communication*, vol. 43, pp. 1–14, 2016.

[53] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3379–3391, 2013.

[54] W. Hachicha, M. Kaaniche, A. Beghdadi, and F. Alaya Cheikh, "No-reference stereo image quality assessment based on joint wavelet decomposition and statistical models," *Signal Processing: Image Communication*, vol. 54, pp. 107–117, 2017.

[55] M. Kaaniche, A. Benazza-Benyahia, B. Pesquet-Popescu, and J.-C. Pesquet, "Vector lifting schemes for stereo image coding," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2463–2475, 2009.

[56] 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.

[57] M. A. Usman and M. G. Martini, "On the suitability of VMAF for quality assessment of medical videos: Medical ultrasound & wireless capsule endoscopy," *Computers in biology and medicine*, vol. 113, p. 103383, 2019.

[58] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 1153–1156.

[59] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[60] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[61] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2009.

[62] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "No-reference video quality assessment using multi-level spatially pooled features," *arXiv preprint arXiv:1912.07966*, 2019.

[63] S. Ahn and S. Lee, "Deep blind video quality assessment based on temporal human perception," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 619–623.

[64] R. Hou, Y. Zhao, Y. Hu, and H. Liu, "No-reference video quality evaluation by a deep transfer cnn architecture," *Signal Processing: Image Communication*, vol. 83, p. 115782, 2020.

[65] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2349–2353.

[66] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[67] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

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

[69] N. Cvejic, A. Loza, D. Bull, and N. Canagarajah, "A similarity metric for assessment of image fusion algorithms," *International journal of signal processing*, vol. 2, no. 3, pp. 178–182, 2005.

[70] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3. IEEE, 2003, pp. III–173.

[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] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on'information measure for performance of image fusion'," *Electronics letters*, vol. 44, no. 18, pp. 1066–1067, 2008.

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

[74] Y. Zheng, E. A. Essock, B. C. Hansen, and A. M. 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.

[75] P.-w. Wang and B. Liu, "A novel image fusion metric based on multi-scale analysis," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on.* IEEE, 2008, pp. 965–968.

[76] H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Information fusion*, vol. 8, no. 2, pp. 193–207, 2007.

[77] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image and vision computing*, vol. 27, no. 10, pp. 1421–1432, 2009.

[78] M. A. Qureshi, A. Beghdadi, and M. Deriche, "Towards the design of a consistent image contrast enhancement evaluation measure," *Signal Processing: Image Communication*, vol. 58, pp. 212–227, 2017.

[79] S.-D. Chen and A. R. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," *IEEE transactions on Consumer Electronics*, vol. 49, no. 4, pp. 1310–1319, 2003.

[80] A. Beghdadi and A. Le Negrate, "Contrast enhancement technique based on local detection of edges," *Computer Vision, Graphics, and Image Processing*, vol. 46, no. 2, pp. 162–174, 1989.

[81] E. Peli, "Contrast in complex images," *JOSA A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[82] M. A. Qureshi, M. Deriche, A. Beghdadi, and M. Mohandes, "An information based framework for performance evaluation of image enhancement methods," in *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA).* IEEE, 2015, pp. 519–523.

[83] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement," *IEEE transactions on cybernetics*, vol. 46, no. 1, pp. 284–297, 2016.

[84] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2387–2390, 2015.

[85] A. Saleem, A. Beghdadi, and B. Boashash, "Image fusion-based contrast enhancement," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 1, p. 10, 2012.

[86] F. Sattar, L. Floreby, G. Salomonsson, and B. Lovstrom, "Image enhancement based on a nonlinear multiscale method," *IEEE transactions on image processing*, vol. 6, no. 6, pp. 888–895, 1997.

[87] A. Chetouani, A. Beghdadi, and M. Deriche, "A new reference-free image quality index for blur estimation in the frequency domain," in *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2009, pp. 155–159.

[88] V. Jaya and R. Gopikakumari, "Iem: a new image enhancement metric for contrast and sharpness measurements," *International Journal of Computer Applications*, vol. 79, no. 9, 2013.

[89] S. S. Agaian, K. Panetta, and A. M. Grigoryan, "Transform-based image enhancement algorithms with performance measure," *IEEE Transactions on Image Processing*, vol. 10, no. 3, pp. 367–382, 2001.

[90] H. Cheng and Y. Zhang, "Detecting of contrast over-enhancement," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 961–964.

[91] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[92] C. Bai and A. R. Reibman, "Controllable image illumination enhancement with an over-enhancement measure," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 385–389.

[93] A. Shokrollahi, A. Mahmoudi-Aznaveh, and B. M.-N. Maybodi, "Image quality assessment for contrast enhancement evaluation," *AEU-International Journal of Electronics and Communications*, vol. 77, pp. 61–66, 2017.

[94] S. Singh and K. Bovis, "An evaluation of contrast enhancement techniques for mammographic breast masses," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 1, pp. 109–119, 2005.

[95] A. Shokrollahi, B. M.-N. Maybodi, and A. Mahmoudi-Aznaveh, "Histogram modification based enhancement along with contrast-changed image quality assessment," *Multimedia Tools and Applications*, pp. 1–22, 2020.

[96] Z. Chen, T. Jiang, and Y. Tian, "Quality assessment for comparing image enhancement algorithms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3003–3010.

[97] T. Saba, A. Rehman, A. Al-Dhelaan, and M. Al-Rodhaan, "Evaluation of current documents image denoising techniques: a comparative study," *Applied Artificial Intelligence*, vol. 28, no. 9, pp. 879–887, 2014.

[98] S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical Signal Processing and Control*, vol. 61, p. 102036, 2020.

[99] P. Coupé, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," *IEEE transactions on image processing*, vol. 18, no. 10, pp. 2221–2229, 2009.

[100] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[101] N. Ponomarenko, S. Krivenko, K. Egiazarian, J. Astola, and V. Lukin, "Weighted mse based metrics for characterization of visual quality of image denoising methods," in *Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM'14)*, 2014.

[102] D. M. Chandler and S. 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.

[103] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on hvs," in *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, vol. 4, 2006.

[104] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Proceedings of the third international workshop on video processing and quality metrics*, vol. 4, 2007.

[105] S. Lu, "No-reference image denoising quality assessment," in *Science and Information Conference*. Springer, 2019, pp. 416–433.

[106] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE transactions on image processing*, vol. 18, no. 4, pp. 717–728, 2009.

[107] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (cpbd)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678–2683, 2011.

[108] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 423–426, 2012.

[109] L. Li, D. Wu, J. Wu, H. Li, W. Lin, and A. C. Kot, "Image sharpness assessment by sparse representation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1085–1097, 2016.

[110] R. Hassen, Z. Wang, and M. M. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2798–2810, 2013.

[111] Q. Sang, H. Qi, X. Wu, C. Li, and A. C. Bovik, "No-reference image blur index based on singular value curve," *Journal of Visual Communication and Image Representation*, vol. 25, no. 7, pp. 1625–1630, 2014.

[112] Y. Liu, J. Wang, S. Cho, A. Finkelstein, and S. Rusinkiewicz, "A no-reference metric for evaluating the quality of motion deblurring." *ACM Trans. Graph.*, vol. 32, no. 6, pp. 175–1, 2013.

[113] K. Zeng, Y. Wang, J. Mao, J. Liu, W. Peng, and N. Chen, "A local metric for defocus blur detection based on cnn feature learning," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2107–2115, 2018.

[114] R. Yan and L. Shao, "Blind image blur estimation via deep learning," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1910–1921, 2016.

[115] S. Yu, F. Jiang, L. Li, and Y. Xie, "Cnn-grnn for image sharpness assessment," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 50–61.

[116] J.-W. Lin, Q. Weng, L.-Y. Xue, X.-R. Cao, and L. Yu, "A retinal image sharpness metric based on histogram of edge width," *Journal of Algorithms & Computational Technology*, vol. 11, no. 3, pp. 292–300, 2017.

[117] V. Simi, D. R. Edla, and J. Joseph, "A fuzzy sharpness metric for magnetic resonance images," *Journal of computational science*, vol. 29, pp. 1–8, 2018.

[118] T. H. Stehle, "Specular reflection removal in endoscopic images," in *Proceedings of the 10th international student conference on electrical engineering.* Citeseer, 2006.

[119] A. Baid, A. Kotwal, R. Bhalodia, S. Merchant, and S. P. Awate, "Joint desmoking, specularity removal, and denoising of laparoscopy images via graphical models and bayesian inference," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017).* IEEE, 2017, pp. 732–736.

[120] G. Anbarjafari, A. Jafari, M. N. S. Jahromi, C. Ozcinar, and H. Demirel, "Image illumination enhancement with an objective no-reference measure of illumination assessment based on gaussian distribution mapping," *Engineering science and technology, an international journal*, vol. 18, no. 4, pp. 696–703, 2015.

[121] S. Salazar-Colores, H. Alberto-Moreno, C. J. Ortiz-Echeverri, and G. Flores, "Desmoking laparoscopy surgery images using an image-to-image translation guided by an embedded dark channel," *arXiv preprint arXiv:2004.08947*, 2020.

[122] C. Wang, A. K. Mohammed, F. A. Cheikh, A. Beghdadi, and O. J. Elle, "Multiscale deep desmoking for laparoscopic surgery," in *Medical Imaging 2019: Image Processing*, vol. 10949. International Society for Optics and Photonics, 2019, p. 109491Y.

[123] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "De-smokegcn: generative cooperative networks for joint surgical smoke detection and removal," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1615–1625, 2019.

[124] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[125] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8160–8168.

[126] N. Clinton, A. Holt, J. Scarborough, L. Yan, P. Gong *et al.*, "Accuracy assessment measures for object-based image segmentation goodness," *Photogramm. Eng. Remote Sens*, vol. 76, no. 3, pp. 289–299, 2010.

[127] W. Huang, N. Li, Z. Lin, G.-B. Huang, W. Zong, J. Zhou, and Y. Duan, "Liver tumor detection and segmentation using kernel-based extreme learning machine," in *Engineering in medicine and biology society (EMBC), 2013 35th annual international conference of the IEEE*. IEEE, 2013, pp. 3662–3665.

[128] R. Cárdenes, R. de Luis-García, and M. Bach-Cuadra, "A multidimensional segmentation evaluation for medical image data," *Computer methods and programs in biomedicine*, vol. 96, no. 2, pp. 108–124, 2009.

[129] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports," *Academic radiology*, vol. 11, no. 2, pp. 178–189, 2004.

[130] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, p. 29, 2015.

[131] J. K. Udupa, V. R. Leblanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn, "A framework for evaluating image segmentation algorithms," *Computerized Medical Imaging and Graphics*, vol. 30, no. 2, pp. 75–87, 2006.

[132] A. Q. Al-Faris, U. K. Ngah, N. A. M. Isa, and I. L. Shuaib, "MRI breast skin-line segmentation and removal using integration method of level set active contour and morphological thinning algorithms," *J Med Sci*, 2013.

[133] G. Gerig, M. Jomier, and C. Valmet, "A new validation tool for assessing and improving 3D object segmentation," in *Proceedings of the International Society and Conference Series on Medical Image Computing and Computer-Assisted Intervention. 2001a*, vol. 1, no. 1, pp. 516–528.

[134] H. Khotanlou, O. Colliot, J. Atif, and I. Bloch, "3D brain tumor segmentation in MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models," *Fuzzy sets and systems*, vol. 160, no. 10, pp. 1457–1473, 2009.

[135] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *computer vision and image understanding*, vol. 110, no. 2, pp. 260–280, 2008.

[136] Z. A. Khan, M. Kaaniche, A. Beghdadi, and F. Alaya Cheikh, "Joint statistical models for no-reference stereoscopic image quality assessment," in *2018 7th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2018, pp. 1–5.

[137] D. Vatolin, A. Bokov, M. Erofeev, and V. Napadovsky, "Trends in S3D-movie quality evaluated on 105 films using 10 metrics," *Electronic Imaging*, vol. 2016, no. 5, pp. 1–10, 2016.

[138] B. Sdiri, A. Beghdadi, F. Alaya Cheikh, M. Pedersen, and O. J. Elle, "An adaptive contrast enhancement method for stereo endoscopic images combining binocular just noticeable difference model and depth information," *Electronic Imaging*, vol. 2016, no. 13, pp. 1–7, 2016.

[139] L. Shen, J. Yang, and Z. Zhang, "Stereo picture quality estimation based on a multiple channel hvs model," in *Image and Signal Processing, 2009. CISP'09. 2nd International Congress on*. IEEE, 2009, pp. 1–4.

[140] P. Gorley and N. Holliman, "Stereoscopic image quality metrics and compression," in *Stereoscopic Displays and Applications XIX*, vol. 6803. International Society for Optics and Photonics, 2008, p. 680305.

[141] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP journal on image and video processing*, vol. 2008, no. 1, p. 659024, 2009.

[142] Z. Zhu and Y. Wang, "Perceptual distortion metric for stereo video quality evaluation," *WSEAS Trans. Signal Process*, vol. 5, no. 7, pp. 241–250, 2009.

[143] A. Maalouf and M.-C. Larabi, "Cyclop: A stereo color image quality assessment metric," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1161–1164.

[144] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, 2013.

[145] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation," *IEEE Transactions on image processing*, vol. 24, no. 5, pp. 1685–1699, 2015.

[146] Z. Sinno, C. Caramanis, and A. C. Bovik, "Towards a closed form second-order natural scene statistics model," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3194–3209, 2018.

[147] Y. Yao, L. Shen, and P. An, "Bivariate analysis of 3D structure for stereoscopic image quality assessment," *Signal Processing: Image Communication*, vol. 65, pp. 128–140, 2018.

[148] W. Miled, J.-C. Pesquet, and M. Parent, "A convex optimization approach for depth estimation under illumination variation," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 813–830, 2009.

[149] G. Verdoolaege and P. Scheunders, "On the geometry of multivariate generalized gaussian models," *Journal of mathematical imaging and vision*, vol. 43, no. 3, pp. 180–193, 2012.

[150] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*, 2010.

[151] J. Yang, C. Hou, Y. Zhou, Z. Zhang, and J. Guo, "Objective quality assessment method of stereo images," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009.* IEEE, 2009, pp. 1–4.

[152] C. T. Hewage and M. G. Martini, "Reduced-reference quality metric for 3D depth map transmission," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010.* IEEE, 2010, pp. 1–4.

[153] R. Akhter, Z. P. Sazzad, Y. Horita, and J. Baltes, "No-reference stereoscopic image quality assessment," in *Stereoscopic Displays and Applications XXI*, vol. 7524. International Society for Optics and Photonics, 2010, p. 75240T.

[154] A. K. Moorthy, C.-C. Su, A. Mittal, and A. C. Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 870–883, 2013.

[155] Z. A. Khan, A. Beghdadi, F. Alaya Cheikh, M. Kaaniche, E. Pelanis, R. Palomar, Å. A. Fretland, B. Edwin, and O. J. Elle, "Towards a video quality assessment based framework for enhancement of laparoscopic videos," in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, vol. 11316. International Society for Optics and Photonics, 2020, p. 113160P.

[156] P. Sanchez-Gonzalez, A. M. Cano, I. Oropesa, F. M. Sanchez-Margallo, F. D. Pozo, P. Lamata, and E. J. Gómez, "Laparoscopic video analysis for training and image-guided surgery," *Minimally Invasive Therapy & Allied Technologies*, vol. 20, no. 6, pp. 311–320, 2011.

[157] S. Bodenstedt, A. Ohnemus, D. Katic, A.-L. Wekerle, M. Wagner, H. Kenngott, B. Müller-Stich, R. Dillmann, and S. Speidel, "Real-time image-based instrument classification for laparoscopic surgery," *Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, 2015.

[158] S. Voros, J.-A. Long, and P. Cinquin, "Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders," *The International Journal of Robotics Research*, vol. 26, no. 11-12, pp. 1173–1190, 2007.

[159] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: a review of the literature," *Medical image analysis*, vol. 35, pp. 633–654, 2017.

[160] J. Zhou and S. Payandeh, "Visual tracking of laparoscopic instruments," *Journal of Automation and Control Engineering Vol*, vol. 2, no. 3, 2014.

[161] S. Bernhardt, S. A. Nicolau, L. Soler, and C. Doignon, "The status of augmented reality in laparoscopic surgery as of 2016," *Medical image analysis*, vol. 37, pp. 66–90, 2017.

[162] M. Siddaiah-Subramanya, M. Nyandowe, and K. W. Tiang, "Technical problems during laparoscopy: a systematic method of troubleshooting for surgeons," *Innovative Surgical Sciences*, vol. 2, no. 4, pp. 233–237, 2017.

[163] E. G. Verdaasdonk, L. P. Stassen, M. van der Elst, T. M. Karsten, and J. Dankelman, "Problems with technical equipment during laparoscopic surgery," *Surgical endoscopy*, vol. 21, no. 2, pp. 275–279, 2007.

[164] A. Chetouani, A. Beghdadi, and M. Deriche, "A hybrid system for distortion classification and image quality evaluation," *Signal Processing: Image Communication*, vol. 27, no. 9, pp. 948–960, 2012.

[165] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[166] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[167] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[168] M. G. Kendall and B. B. Smith, "On the method of paired comparisons," *Biometrika*, vol. 31, no. 3/4, pp. 324–345, 1940.

[169] A. Leibetseder, M. J. Primus, S. Petscharnig, and K. Schoeffmann, "Real-time image-based smoke detection in endoscopic videos," in *Proceedings of the on Thematic Workshops of ACM Multimedia*, 2017, pp. 296–304.

[170] J. Immerkaer, "Fast noise variance estimation," *Computer vision and image understanding*, vol. 64, no. 2, pp. 300–302, 1996.

[171] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[172] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[173] Z. A. Khan, A. Beghdadi, M. Kaaniche, and F. Alaya Cheikh, "Residual networks based distortion classification and ranking for laparoscopic image quality assessment," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 176–180.

[174] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2002, pp. IV–3313.

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

[176] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE transactions on broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[177] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," *arXiv preprint arXiv:2002.10651*, 2020.

[178] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.

[179] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3773–3777.

[180] L. Xu, J. Li, W. Lin, Y. Zhang, L. Ma, Y. Fang, and Y. Yan, "Multi-task rank learning for image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1833–1843, 2016.

[181] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks." in *ACM Multimedia*, 2018, pp. 546–554.

[182] D. Varga, "No-reference video quality assessment based on the temporal pooling of deep features," *Neural Processing Letters*, vol. 50, no. 3, pp. 2595–2608, 2019.

[183] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *arXiv preprint arXiv:2005.14354*, 2020.

[184] X. He, K.and Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[185] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[186] T.-S. Nguyen, L. Ngo, M. Luong, M. Kaaniche, and A. Beghdadi, "Convolution autoencoder based sparse representation wavelet for image classification," in *IEEE Workshop on Multimedia and Signal Processing (MMSP)*, 2020, pp. 1–6.

[187] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–709.

[188] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[189] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3311–3319.

[190] F. Götz-Hahn, V. Hosu, and D. Saupe, "Comment on" no-reference video quality assessment based on the temporal pooling of deep features"," *arXiv preprint arXiv:2005.04400*, 2020.

[191] Z. Khan, A. Beghdadi, F. Cheikh, M. Kaaniche, and M. Qureshi, "A multi-criteria contrast enhancement evaluation measure using wavelet decomposition," in *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020.

[192] R. Naseem, F. Alaya Cheikh, A. Beghdadi, O. J. Elle, and F. Lindseth, "Cross modality guided liver image enhancement of CT using MRI," in *2019 8th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2019, pp. 46–51.

[193] J. Liu, C. Zhou, P. Chen, and C. Kang, "An efficient contrast enhancement method for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1715–1719, 2017.

[194] B. Singh, R. S. Mishra, and P. Gour, "Analysis of contrast enhancement techniques for underwater image," *International Journal of Computer Technology and Electronics Engineering*, vol. 1, no. 2, pp. 190–194, 2011.

[195] M. E. Celebi, M. Lecca, and B. Smolka, *Color Image and Video Enhancement*. Springer, 2015, vol. 4.

[196] A. Beghdadi, M. A. Qureshi, B. Sdiri, M. Deriche, and F. Alaya-Cheikh, "Ceed-a database for image contrast enhancement evaluation," in *2018 Colour and Visual Computing Symposium (CVCS)*. IEEE, 2018, pp. 1–6.

[197] S. A. Amirshahi, A. Kadyrova, and M. Pedersen, "How do image quality metrics perform on contrast enhanced images?" in *2019 8th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2019, pp. 232–237.

[198] S. S. Agaian, B. Silver, and K. A. Panetta, "Transform coefficient histogram-based image enhancement algorithms using contrast entropy," *IEEE transactions on image processing*, vol. 16, no. 3, pp. 741–758, 2007.

[199] A. Beghdadi, M. A. Qureshi, and M. Deriche, "A critical look to some contrast enhancement evaluation measures," in *2015 Colour and Visual Computing Symposium (CVCS)*. IEEE, 2015, pp. 1–6.

[200] S. Mukhopadhyay and B. Chanda, "A multiscale morphological approach to local contrast enhancement," *Signal Processing*, vol. 80, no. 4, pp. 685–696, 2000.

[201] R. Hummel, "Image enhancement by histogram transformation," *Computer Graphics and Image Processing*, vol. 6, no. 2, pp. 184–195, 1977.

[202] T. Briand, J. Vacher, B. Galerne, and J. Rabin, "The heeger & bergen pyramid based texture synthesis algorithm," *Image processing on line*, vol. 4, pp. 276–299, 2014.

[203] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proceedings., International Conference on Image Processing*, vol. 3. IEEE, 1995, pp. 444–447.

[204] Y. Zhang, B. Jiang, J. Wu, D. Ji, Y. Liu, Y. Chen, E. X. Wu, and X. Tang, "Deep learning initialized and gradient enhanced level-set based segmentation for liver tumor from CT images," *IEEE Access*, 2020.

[205] S. C.-F. Lin, C. Y. Wong, M. A. Rahman, G. Jiang, S. Liu, N. Kwok, H. Shi, Y.-H. Yu, and T. Wu, "Image enhancement using the averaging histogram equalization (avheq) approach for contrast improvement and brightness preservation," *Computers & Electrical Engineering*, vol. 46, pp. 356–370, 2015.

[206] C. Y. Wong, S. Liu, S. C. Liu, M. A. Rahman, S. C.-F. Lin, G. Jiang, N. Kwok, and H. Shi, "Image contrast enhancement using histogram equalization with maximum intensity coverage," *Journal of Modern Optics*, vol. 63, no. 16, pp. 1618–1629, 2016.

[207] H. Ibrahim and N. S. P. Kong, "Brightness preserving dynamic histogram equalization for image contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 4, pp. 1752–1758, 2007.

[208] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 593–600, 2007.

[209] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[210] Z. Al-Ameen and G. Sulong, "A new algorithm for improving the low contrast of computed tomography images using tuned brightness controlled single-scale retinex," *Scanning*, vol. 37, no. 2, pp. 116–125, 2015.

[211] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[212] G.-C. Lin, W.-J. Wang, C.-C. Kang, and C.-M. Wang, "Multispectral MR images segmentation based on fuzzy knowledge and modified seeded region growing," *Magnetic resonance imaging*, vol. 30, no. 2, pp. 230–246, 2012.

[213] D. Dreizin, U. K. Bodanapally, N. Neerchal, N. Tirada, M. Patlas, and E. Herskovits, "Volumetric analysis of pelvic hematomas after blunt trauma using semi-automated seeded region growing segmentation: a method validation study," *Abdominal Radiology*, vol. 41, no. 11, pp. 2203–2208, 2016.

[214] J. Lian, Y. Ma, Y. Ma, B. Shi, J. Liu, Z. Yang, and Y. Guo, "Automatic gallbladder and gallstone regions segmentation in ultrasound image," *International journal of computer assisted radiology and surgery*, vol. 12, no. 4, pp. 553–568, 2017.

[215] N. Satpute, R. Naseem, E. Pelanis, J. Gómez-Luna, F. Alaya Cheikh, O. J. Elle, and J. Olivares, "Gpu acceleration of liver enhancement for tumor segmentation," *Computer Methods and Programs in Biomedicine*, vol. 184, p. 105285, 2020.

[216] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[217] J. Mukherjee and S. K. Mitra, "Enhancement of color images by scaling the DCT coefficients," *IEEE Transactions on Image processing*, vol. 17, no. 10, pp. 1783–1794, 2008.

[218] S. Chen and A. Beghdadi, "Natural enhancement of color image," *EURASIP Journal on Image and Video Processing*, vol. 2010, no. 1, pp. 1–19, 2010.

[219] N. Hassan and N. Akamatsu, "A new approach for contrast enhancement using sigmoid function," *The International Arab Journal of Information Technology*, vol. 46, no. 2, pp. 221–226, 2004.

[220] S. Chen, A. Beghdadi, and M. Cheriet, "Degraded color document image enhancement based on nrcir," in *2010 2nd European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2010, pp. 19–22.

[221] N. Al-Najdawi, M. Biltawi, and S. Tedmori, "Mammogram image visual enhancement, mass segmentation and classification," *Applied Soft Computing*, vol. 35, pp. 175–185, 2015.

[222] T. Hopp, M. Dietzel, P. A. Baltzer, P. Kreisel, W. A. Kaiser, H. Gemmeke, and N. V. Ruiter, "Automatic multimodal 2D/3D breast image registration using biomechanical fem models and intensity-based optimization," *Medical image analysis*, vol. 17, no. 2, pp. 209–218, 2013.

[223] M. A. Qureshi, M. A. Deriche, A. Beghdadi, and M. Mohandes, "An information based framework for performance evaluation of image enhancement methods," in *2015 International Conference on Image Processing Theory, Tools and Applications, IPTA 2015, Orleans, France, November 10-13, 2015*, 2015, pp. 519–523.

[224] I. France, "3Dircadb, 3D image reconstruction for comparison of algorithm database," 2016.