

**Improving methods for perceptual image quality
assessment.**

by

Navaneeth Kamballur Kottayil

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

© Navaneeth Kamballur Kottayil, 2018

Abstract

Image quality assessment (IQA) algorithms aim to simulate human judgment of visual quality on an image. These algorithms are essential components of every multimedia pipeline. IQA is divided into full reference(FR) or no reference (NR) depending on the presence or absence of a pristine image while the image is being judged. Traditional FR and NR-IQA algorithms show high performance on conventional images. However, they fail to generate good results on newer modalities of multimedia data like HDR images and 3D textured meshes. Furthermore, these IQA algorithms limit themselves to low-level features and do not incorporate the effect of image content on human judgment of visual quality. In this thesis, we explore and extend the current IQA capabilities to address these issues.

We focus on adaptation of IQA to newer modalities of multimedia content, incorporation of scene level knowledge and integration of low-level features. The goal of this thesis is to advance the IQA algorithms to perform on a larger range of multimedia content by accounting for all available data features and applying latest Machine Learning technologies.

The new modalities of multimedia content that we explore in this thesis are High Dynamic Range (HDR) images, and 3D textured meshes (tex-meshes).

HDR images pose challenges to conventional IQA methods because of the much larger range of perceptual effects shown by them. This leads to the failure of statistics-based approaches followed by NR-IQA techniques. To account for perceptual effects of HDR, we designed machine learning models that can

determine human visual sensitivity from subjective image quality data, without going through psycho-visual experiments. Using this model, we developed a blind noise estimation and quality assessment algorithm for HDR images.

Next, we addressed the lack of research into the perceptual effects of texture in tex-meshes. We performed subjective experiments to model the effects of texture compression and incorporated our results into existing research into 3D mesh quality assessment.

Furthermore, we design two algorithms that show better correlation to human judgments on quality compared to the existing FR-IQA. The first algorithm is a content-specific IQA performance enhancer, which can be applied to any IQA. The second algorithm is a new full reference algorithm that integrates more low-level features and color elements to improve IQA accuracy.

Finally, we performed a case study that analyzed the changes in gaze behavior of humans with the level of familiarity of task. We show statistically significant differences in gaze behavior dependent on familiarity.

We validate all of our proposed algorithms by comparing the predictions of our algorithms with human opinions. We observed a high degree of correlation between the human and algorithms scores.

Preface

The majority of the contents of this thesis has been published or are under review in peer reviewed journals and conferences. The contents of Chapter 3 presents the first work in no reference HDR image quality assessment and has been published in *IEEE Transactions of Image Processing*. Chapter 4 details our method of deriving perceptual error thresholds from IQA databases and is under review in *IEEE International Conference on Image Processing*. Chapter 5 is the first effort in the literature to incorporate a true scene dependent processing into an IQA and has been published in *IEEE systems Man and Cybernetics*. Chapter 6 provides our results that show improvements to state of the art in conventional full reference IQA with low level feature integration and has been published in *Springer - Signal, Image and Video Processing*. Chapter 7 is the result of our study into gaze behavior in the surgical environment and has been published in *IEEE Engineering in Medicine and Biology Society and American Journal of Innovative Research and Applied Sciences*. Chapter 8 is on study of effect of texture compression on 3D textured mesh and is under review in *International Conference on Smart Multimedia*. I choose to use first-person plural throughout this thesis to honor the contributions of my advisors and collaborators on my various works.

Ethics Approval

The subjective experiments in this was covered by Ethics Approval # Pro00016136 of the University of Alberta.

Acknowledgements

Firstly, I would like to express my heart felt gratitude to my advisors Dr. Anup Basu and Dr. Irene Cheng for their continuous support and guidance. Thank you for your patience, motivation, enthusiasm, and immense knowledge that helped me grow as a researcher. I could not have asked for better mentors.

Secondly, I am grateful to Dr. Frederic Dufaux, Dr. Giuseppe Valenzise, Dr. Guan-Ming Su and Parwant Ghuman, for offering me with internship opportunities in their groups/companies. This resulted in me working on diverse and exciting projects. I thank the research teams at Dolby Laboratories and 3VGeomatics for their mentorship and support.

I appreciate the collaborative efforts of my colleagues Subhayan Mukherjee, Xinyao Sun at the University of Alberta, and Emin Zerman at Telecom Paris-technic. I am also thankful to my friends at the University of Alberta: Housam, Nathaniel, Nasim, Bhaskar, Rohit, Yathirajan, Samreen, Faisal, Vivek, Gautham, Ankush, Ujjwal, Ziyoun, Derek, and Winny for all the fun discussions on life and research over coffee. I am also thankful to my friends at Telecom Paris-technic for the stimulating discussions, and the adventures we had in Paris.

I would like to thank my thesis committee: Dr. Nilanjan Ray, Dr. Pierre Boulanger, and Dr. Bruce Cockburn, for their encouragement, insightful comments, and thoughtful questions, which helped me become a better scientist.

I would like to thank my family Padmanabhan P. T., Asha K. K., Navajoth K. K., Sukesh Nambiar and Dr. Murlidharan E. K., Megha Panda and the rest of my family, for always being there for me.

Lastly, I would like to thank each and everyone who were directly and indirectly were a part of my Ph.D. journey, without your support this thesis would not be possible.

Contents

1	Introduction	1
1.1	Scope and Significance	3
1.1.1	Application	3
1.1.2	Open problems	3
1.2	Motivation	4
1.2.1	Dynamic Range	4
1.2.2	Content dependency	5
1.2.3	Low-level color feature integration	5
1.2.4	Quality of textured 3D mesh	6
1.3	Contributions	7
1.4	Challenges	8
1.5	Organization of the thesis	9
2	Background and Related Work	11
2.1	2D image quality	11
2.1.1	Full Reference Image quality assessment	11
2.1.2	No reference Image Quality assessment	17
2.2	High-level features in Image quality assessment	19
2.2.1	Visual saliency	19
2.2.2	Saliency and image quality	20
2.2.3	Image aesthetics and high-level features	21
2.3	HDR image quality	21
2.3.1	Tone Mapped IQA	22
2.3.2	Algorithms for comparing performance of HDR NR-IQA	23
2.4	Perceptual Quality of 3D textured meshes	23
2.5	Datasets used for evaluation of results	24
2.5.1	Low dynamic range datasets	25
2.5.2	High Dynamic Range datasets	25
3	Quality assessment of High Dynamic Range Images	27
3.1	Introduction	27
3.2	Motivation	29
3.2.1	Perceptual factors affecting HDR data	29
3.3	Proposed method	30
3.3.1	Design	30
3.3.2	Error Estimation	31
3.3.3	Perceptual Resistance	32
3.3.4	Mixing Function	33
3.3.5	Training	36
3.3.6	Further considerations	38
3.4	Architecture	38
3.5	Results	40
3.5.1	Dataset	40

3.5.2	Reference IQA schemes	40
3.5.3	Learning performance	42
3.5.4	Performance comparisons	43
3.5.5	Generalization capability	46
3.5.6	Error Estimation	47
3.5.7	Perceptual Resistance	48
3.5.8	Error maps	52
3.5.9	Effects of mixing function	53
3.5.10	Failure cases	55
3.6	Conclusion	56
4	Learning error visibility from Image quality	57
4.1	Introduction	57
4.2	Motivation	58
4.3	Proposed model	59
4.3.1	Mathematical framework and assumptions	59
4.3.2	Implementation	60
4.4	Results and Analysis	62
4.4.1	Performance	62
4.5	Conclusion	65
5	Content dependency	67
5.1	Introduction	67
5.1.1	Psycho-visual experimental evidence	68
5.1.2	Experiment on CSIQ	69
5.2	Computational Design and Implementation	69
5.2.1	Image Content C	70
5.2.2	Image Features F	71
5.2.3	Problem formulation	72
5.3	Results and Analysis	75
5.3.1	Visual Error Importance (VEI) maps	77
5.3.2	Limitation and Future direction	78
5.4	Conclusion	78
6	Color and low-level-feature based quality assessment	79
6.1	Introduction	79
6.2	Computational Model	80
6.2.1	Color adaptation	84
6.3	Experimental results	86
6.3.1	Performance comparison and analysis	86
7	Investigation of Gaze Patterns in a case study	92
7.1	Introduction	92
7.2	Motivation	93
7.3	Method	93
7.3.1	Subjects and experimental environment	94
7.3.2	Tasks	94
7.4	Analysis	95
7.4.1	Identification of experts using time performance	96
7.4.2	Features used	97
7.5	Conclusion	100

8	Impact of JPEG compression on perceptual quality of 3D models	102
8.1	Introduction	102
8.2	Experimental setup	103
8.2.1	Design consideration	103
8.2.2	Interface	104
8.2.3	Experimental conditions	105
8.2.4	Stimuli generation	105
8.2.5	Experiment	105
8.3	Mathematical Modeling	106
8.3.1	Analysis	107
8.3.2	Future work	109
8.4	Conclusion	110
9	Conclusions and future directions	111
	References	113

List of Tables

2.1	LDR database statistics.	26
3.1	Database statistics.	41
3.2	Overall Performance comparison.	45
3.3	Statistical significance of difference in performance of algorithms (Bolded in table 3.2).	46
3.4	Generalization capability.	49
3.5	Performance with various mixing functions.	55
4.1	Performance comparison between different algorithms. Highlighted are the best state-of-the-art methods (handcrafted and CNN-based), as well as our results.	65
5.1	Performance of IQMs before and after augmentation with VEI.	75
6.1	Comparison of performance on datasets.	86
6.2	Performance of using a single window compared with using a center-surround analysis.	90
8.1	Details of 3D models used for psychovisual experiment.	105

List of Figures

1.1	Schematic representation of Full Reference IQA.	2
1.2	Schematic representation of No Reference IQA.	3
2.1	Example images in the LIVE dataset.	26
3.1	Proposed strategy for IQA. E-net detects the error, P-net detects the perceptual response, Mixing function combines the results to produce a DMOS.	31
3.2	Behavior of the mixing function. (A) Varying T with k fixed at 1. (B) Varying k with T fixed at 20.	35
3.3	Two-stage training Process.	37
3.4	Network structure for Error estimation.	39
3.5	Network structure for Perceptual Resistance value.	39
3.6	Training error vs batch number for every image in randomly sampled training set (570 images).	42
3.7	SRCC performance after training with increasing number of samples.	43
3.8	Scatter plot between objective scores by proposed method and actual DMOS.	48
3.9	Left column distorted images, second column ground truth error, third column the error estimation results from E-net and final column shows the MSE between Estimated and Actual errors on the HDR luminance image.	50
3.10	Image for studying the nature of Perceptual Resistance.	50
3.11	Input image (top row), Estimated error by E-net (second row), Perceptual Resistance by P-net (third row) and the local error map from the mixing function (fourth row) for different luminance ranges obtained by linearly scaling image intensities in range $[0,4000]$. Red implies high value and blue implies low.	51
3.12	Comparison of distortions in image or error maps estimated by various IQA schemes. Red represents high value and blue represents a low value.	52
3.13	Comparison of the output of P-net with HDR-VDP2.2 contrast threshold. Image pixel values are in log scale and normalized.	52
3.14	Comparisons of results of mixing function (Error map) and P-net (Perceptual Resistance) when different mixing functions are used for training the system. HDR-VDP probability of error detection and contrast thresholds are shown in first column for reference.	54
3.15	Cases where the error maps produced by proposed method fails.	55
4.1	Neural network architecture for extracting contrast detection thresholds from IQA databases.	60

4.2	Scatter plots of contrast detection thresholds derived using our method vs experimentally measured values. The system is trained on (a) LIVE dataset (b) TID 2013 dataset. The polynomial fitting line is show in red. Red points correspond to patches whose luminance is outside the range of the training datasets, see Figure 4.3. r is the PLCC (after fitting) on the whole test set; r' is the PLCC excluding the red points.	63
4.3	Distribution of average intensity of 32×32 patches in LIVE, CSIQ and TID datasets. The distributions of the three datasets overlap only in the intensity range $[10, 250]$	64
4.4	Examples of local visibility thresholds produced by our method trained on TID 2013. The estimated thresholds are decimated to match the resolution of the ground truth. Both sets of values are scaled to fit in the range $[0, 1]$ for visualization, with 0 (black) being the lowest threshold.	66
4.5	Some failure cases of the proposed method trained on TID 2013. Often these cases correspond to patches with low luminance, which are underrepresented in the training data.	66
5.1	Examples of outdoor natural and outdoor man-made images chosen from the CSIQ dataset.	70
5.2	Scene Probability: The description on top shows the likelihood of each class in the image: Id refers to indoor, Od-Nat is outdoor natural, Od-Man is outdoor man-made.	71
5.3	Visual presentation of the feature maps extracted from an image. The colorbar with the relative values from cool to hot are shown next to the image. The original image is shown on top left. The Visual Error Importance (VEI) map is shown on the right.	72
5.4	Schematic representation of the proposed strategy.	73
5.5	We compare the VEI map for the correct class, with the VEI maps generated by forcing the algorithm to consider the image as a different class. Class is shown on the left. The values are normalized to $[0,1]$ to show relative importance. The colorbar on the right of the image shows increasing visual error importance (cool to hot colors).	77
6.1	Outputs at various stages (a) The original image, (b) Three filter outputs, (c) Results after blockwise normalization, (d) Schematic representation of frequency scaling, (e) Pooling, (f) Pictorial representation of the final feature.	80
6.2	Example images illustrate the general trend of the coefficient variance associated with a feature map. It decreases when moving across levels.	83
6.3	The consolidated result from thirty images also show that the variance coefficients in feature maps decrease when moving across levels.	83
6.4	An optimization formulation can be obtained by adjusting the values of K_1 and K_2	84
6.5	Colorful images (a),(b) and (c) are recognized by $Cr \geq 0.25$. Example images with $Cr < 0.25$ are shown in (d),(e) and (f).	84
6.6	Scatter plots of our metric scores vs subjective scores on dataset (a) CSIQ, (b) TID, (c) LIVE and (d) TID2013.	90
7.1	2D camera positions in training box.	95

7.2	Three views.	95
7.3	Task that the subjects were asked to complete.	96
7.4	Object detection and tracking; the green spots denote the tracked points, the red spot denotes the target point. The figure on the right shows a segmented view of the pixels being tracked.	96
7.5	Number of times the Gaze of the user shifted between top and front view by the high and low performers.	98
7.6	Percentage of Top and Front view used by the high and the low performers.	98
7.7	Front/Top ratio for high and low group performers.	99
7.8	High performer. Here red plots are the cases where the top view was used. The blue are the cases where the front view was used.	99
7.9	Low performer. Here red plots are the cases where the top view was used. The blue are the cases where the front view was used.	100
8.1	Screen shot of the interface used for rating based experiment.	104
8.2	Visualization of models used in experiment.	106
8.3	Average score of users for all models vs JPEG Q factors.	107
8.4	Change in perceived quality for each Q factor.	108
8.5	Change in perceived quality for different texture file sizes.	109
8.6	Average user score for each model at different levels of compression.	109

Chapter 1

Introduction

An ever increasing focus on multimedia content creation in the modern world has led to the rapid development of high quality 2D and 3D imaging devices. The multimedia processing pipeline has various operations, which include content acquisition, storage, compression, transmission and visualization. Some practical examples of these operations are JPEG compression, contrast changes, color adjustments for images etc. In order to optimize the limited resources such as time and bandwidth, these processing operations can introduce visual artifacts to the images. Blurring or lose of contrast due to image compression is a common example; these changes usually reduce the visual quality of the content.

An objective measure of the visual quality change is essential in understanding and fine tuning of the various processing algorithms and for comparisons of the final results of the processing operations. Since the target audience of any multimedia system is the human visual system (HVS), the method used to evaluate these processing operations is through a user study, with human subjects scoring the changes on a fixed scale using a wide variety of subjective experiments. The test subjects rate the images on an intuitive scale like excellent, best, good and worst. These scales are then transferred to numerical values after statistical normalization operations [37], [58]. These numerical values indicate the perceived quality of the images by the HVS.

Broadly speaking, experimental methods to measure the quality of a content can be classified into the categories of full reference and no-reference/reduced

reference. In full reference perceptual quality assessment, an altered/processed content is compared with an unaltered ground truth content to produce a relative quality difference score (Fig 1.1). In no-reference, the content is evaluated independently to determine a *quality score* for the content (Fig 1.2).

The subjective experiments for objectively measuring perceived quality are tedious, time consuming and not viable for testing on a large database of thousands of images. To solve this problem, a category of automated algorithms called as perceptual quality assessment algorithms or perceptual quality metrics are used. These algorithms give a quality score by comparing the multimedia content using mathematical models of HVS. Estimating perceptual quality is one of the most challenging problems in the field of image processing. This is because to be successful, the algorithm's mathematical model needs to be similar to the behavior of HVS. Perceptual quality metrics are often integrated with imaging algorithms to achieve the best compromise between efficiency and perceptual quality. A classical example is image or video compression, but the metrics have been also used in graphics to control global illumination solutions or find the optimal tone-mapping curve.

Perceptual quality assessment algorithms can be divided into full reference and no-reference, depending on the full reference or no-reference results it tries to predict. In case of full reference methods, the algorithm will need access to both the undistorted reference and the distorted target.

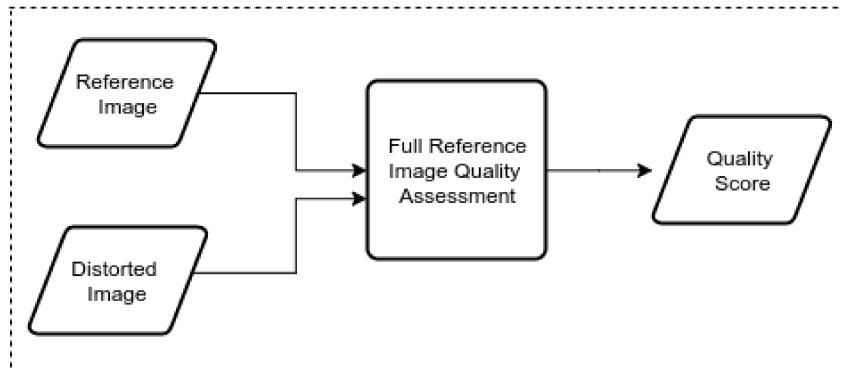


Figure 1.1: Schematic representation of Full Reference IQA.

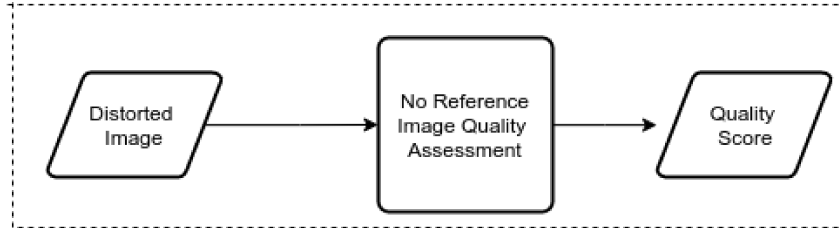


Figure 1.2: Schematic representation of No Reference IQA.

1.1 Scope and Significance

Image quality assessment is a vast and mature field. The image quality assessment (IQA) techniques covered under the scope of this thesis, studies the impact of low-level distortions like additive white Gaussian noise (AWGN), compression artifacts, and localized distortions. We do not address the aesthetic changes to the image like a change in composition, lighting, positioning of a subject in a photograph etc.

1.1.1 Application

IQA algorithms play an important role in the video and image processing pipeline and are commonly used to best optimize visual quality for any given set of variables. This helps in removing the need to perform expensive user studies to validate different methods of image processing. In addition to this, a good model may be used to study the impact of algorithms on data range beyond the conventional values. This is useful in design of new displays, processing algorithms etc. A classical example is image or video compression, but the metrics have been also used in graphics to control global illumination solutions [92], or find the optimal tone-mapping curve [7].

1.1.2 Open problems

New image modalities like high dynamic range imaging is showing the limitations of existing methods and there is much demand in the industry to overcome these limitations. Some notable issues here are the absence of no reference models that are perceptual, lack of easily usable models etc.

Additionally, areas of research into IQA have ignored certain aspects of images like the color and content. These areas also need to be explored to model a comprehensive method of assessing visual quality.

In this thesis, we focus on three aspects of assessing visual quality of images, namely Modalities (Dynamic range - Chapter 3 , 3D - Chapter 8), Content (Chapter 4) and Low level color feature integration (Visibility - Chapter 4, Feature integration - Chapter 6, Expert gaze analysis -7). Within these sub-fields, we investigate full reference methods for color and content, no reference methods for dynamic range and texture compression in 3D textured meshes. We choose these topics of study to bridge the gap in these sub-fields of IQA.

1.2 Motivation

Perceptual Quality assessment is a heavily researched field with a lot of improvements over the years. However, there are some aspects of modeling visual quality that have not yet been adequately studied. In this thesis, we focus on three aspects of perceptual quality assessment: Dynamic range, Content dependency, and low level feature integration.

1.2.1 Dynamic Range

High dynamic range (HDR) imaging can capture a much larger range of luminance compared to conventional images. This larger range of luminance is achieved by increasing the number of bits used to store the image pixels. Typically HDR image content is stored using 16-32 bit floating point values per pixel as opposed to a conventional image file that uses 8 bits per integer for a pixel. HDR imaging technology incorporates techniques of capturing, processing and displaying HDR content. This technology is becoming mainstream in the consumer market and is reaching larger number of people as a result of investments in the technology made by the camera and television manufacturers. IQA of the commonly seen image distortions on HDR images is a problem of great academic and industrial interest. Existing IQA metrics developed for LDR fail to provide adequate performance in measurement of

visual quality on HDR data. Hence, there is a need for high performance automated systems that are capable of IQA. Though there are some stellar works on full reference high dynamic range images, at the time of writing, there is no published work on NR-IQA for HDR images that have been evaluated by users on HDR compatible screens.

1.2.2 Content dependency

In the HVS, the quality evaluation process is carried out at different levels starting from the pixel-level to the scene or content level. An explanation of how high-level scene information affects low-level processing can be found in [34]. The authors showed that low-level feature maps can be preset at a higher-level and influence in a top-down manner. In other words, viewer expectation can have a certain degree of influence when interpreting low-level features. Thus, depending on the viewer's expectation of the scene, his/her visual quality score maybe subject to change. Another major experimental evidence was provided by [101]. The study analyzed influence of scene categories on image visual annoyance. Specifically they selected: indoor, outdoor natural, and outdoor man-made. The study reported *statistically significant results showing that a scene category had an influence on the degree of visual annoyance* perceived by the viewer. Specifically, they found that people are more critical to indoor images as compared with outdoor images, especially when comparing with outdoor natural scenes. Another observation in the experiment was the difference in quality assessment, which seems dependent on scene category. Though there are significant evidence of influence of content on visual quality, we found inadequate research on modeling this influence and incorporating this into the visual quality assessment algorithms.

1.2.3 Low-level color feature integration

Almost all of the existing image quality metrics focus on the luminance of the pixel as a measure of quality and ignore the color channels completely. However, color is a significant part of the visual perception process and needs to be considered while evaluating the perceived quality of images. It is believed

that sensitivity of the HVS to scene content is derived from various stimulus modalities, including intensity, color, spatial and temporal (for dynamic scene only) features. This is consistent with what we noticed from our datasets, where brightly colored images display a different perceptual sensitivity compared to the less vibrant images. Additionally, the expertise of the user could also have an influence of the image analysis process. Here also research is lacking. The effect of expertise can be studied by analyzing the eye gaze behavior of an expert and a novice. We can gain additional insights by the statistical analysis of the data collected. Hence there is room for integration of color into IQA metrics.

1.2.4 Quality of textured 3D mesh

With the advent of VR and computer games, 3D graphics is becoming more relevant. Textured polygonal meshes (referred to as *tex-meshes* in this thesis) are an integral part of 3D graphics. Most tex-mesh representations make use of uncompressed texture formats for the sake of preserving quality. Hence, most applications have large texture sizes consuming large file size and bandwidth during transmission. A good analysis of what texture resolution to use vs. quality of the mesh given a limited bandwidth is given in [87]. The study, however, did not look into the impact of texture compression.

JPEG is one of the most popular image compression methods to date and is widely used. Most applications involving subjective judgments on quality use a fixed compression ratio that is heuristically decided. For JPEG, the compression ratio is decided by the Q factor used in the compression of the images. There are however no guidelines on how to select this Q factor. Determining the ideal compression could be greatly beneficial towards transmission and optimization of 3D models. We found no other studies exploring the effect of texture compression on 3D meshes though there are a lot of unsolved research problems in this area.

1.3 Contributions

The following are the contributions of the thesis.

- *Introduction of a new method to learn perceptual visibility from Image Quality datasets.* As opposed to the existing methods of perceptual sensitivity measurements with psychovisual experiments, we provide a data driven solution to this problem. We use a neural network based architecture to derive perceptual sensitivity measure from a real world image quality database. Specifically, we define a new term called 'Perceptual Resistance' of a patch in an image. This represents how difficult it is for a viewer to perceive the average visual error present in the block. This can be combined with any error detection based method to provide the perceived quality estimate of any high dynamic range image.
- *Proposal of the state-of-the-art No reference assessment of compression artifacts in HDR images.* We provide a neural network based architecture for the detection of compression artifacts in high dynamic range images without the need for a reference image. We use perceptual resistance that we derived to predict the perceived visual quality of HDR images. This was validated on multiple datasets of HDR quality measurements and was found to be the state-of-the-art in the field.
- *Modeling the influence of scene content on full reference assessment of visual quality.* We develop a new architecture for improving full reference visual quality assessment in images. Our low-level feature generation approach based on a single-layer object detection architecture is especially effective for IQA because of the following advantages: 1) the generated features capture structural information [108] of the image, 2) our approach, with a new frequency scaling technique, we can explicitly model the HVS and account for the first mechanism to detect not easily visible distortions [57], 3) our approach captures the low-level features that can be used for object detection accounting for the second mechanism to detect supra-threshold distortions [57], 4) the extracted

low-level features can influence visual perception based on earlier findings [35], 5) visual saliency is incorporated in our framework by using our new center-surround processing step.

- *Investigation on gaze patterns, using a surgical environment in our case study* We performed a user study analyzing the gaze patterns of a high performing experts in a multi-view laparoscopic environment and compared it with that of a novice to detect the differences between the gaze behavior. Our finding leads to a future research direction of analyzing the difference in IQA between viewers with and without experience of the scene or image content.
- *Investigation into impact of texture compression on perceptual quality of 3D textured mesh* We study the effect of JPEG texture compression on the perceived quality of a 3D mesh. Our experiment provides a statistics-based model for describing the drop in quality of texture with decrease in Q factor. We integrated these results into current research.

1.4 Challenges

Each of the distinct avenues of investigation has its own set of challenges that needs to be handled

- *High dynamic range images on assessment of visual quality.* The main challenge in HDR images is the lack of datasets and the expense of conducting user study. The use of publicly available datasets for LDR images is not suitable because of the use of LDR based screens in user studies. At the time of writing, there are only a few publications that studied the effects of distortions on HDR quality. These were individual studies that were conducted on a single type of distortion and did not cover the impact of multiple distortions at the same time. The second challenge here is the lack of localized perceptual data. Most conventional methods of deriving perceptual data involved experiments with simple stimuli that cannot generalize to real world complex images. We face

the problem of modeling these perceptual effects without adequate data on interactions between the effects.

- *Study of influence of image content on visual quality.* There is some literature on how scene content can influence visual quality but the effects of this have never been modeled explicitly. The easiest solution to the problem is to divide an image based on content and then assess the qualities of different content types separately. However, the diversity of content in image quality dataset are not large enough and there is not enough examples of each content type. This makes it difficult to model the effect.
- *Investigation into features used in image quality assessment scheme* Image quality feature and its integration is an extensively studied problem but there is no consistent approach by which we can integrate this theory. There is also the problem of modeling color and its impact on HVS which have largely been ignore in quality assessment literature. Different algorithms seem to show different performances on different datasets and this makes it hard to quantify how well a method works.
- *Texture mesh quality interaction in 3D models* The major hurdle in the study of 3D perceptual quality is the lack of sufficient data to model the phenomenon. Most user studies that are publicly available concentrate on LDR images.

1.5 Organization of the thesis

The rest of the thesis document is organized as follows: In Chapter 2, we cover the background research and related works in the areas of image quality assessment. In Chapter 3 and 4, we detail our study of dynamic range and its effect on image quality. A no reference image quality assessment model on image quality is presented as the application of the observations from our study. Chapter 5 covers our findings on content dependency in image quality

assessment and the proposals to augment image quality metrics with this result. In Chapter 6 we list our approach in modeling color into full reference image quality assessment strategies. In chapter 7 we present the results of a study we did to apply the results we derived in color features to the field of surgical performance enhancement. In Chapter 8, we report our results of the user study on 3D quality and mathematical modeling. Finally, in Chapter 9 we state our conclusions and indicate directions for future work.

Chapter 2

Background and Related Work

There is an extensive amount of research that has gone into quality assessment research. We conceptually separate the topics of interest in the following order and discuss it in the following sections.

1. 2D Image Quality
2. High level feature based image quality
3. HDR image quality
4. Quality of textured 3D mesh
5. Datasets

2.1 2D image quality

The research in 2D image IQA can be broadly classified into full reference and no reference methods.

2.1.1 Full Reference Image quality assessment

The most simple way of comparing the quality of two images is via peak signal-to-noise-ratio (PSNR), mean squared error (MSE) etc.

PSNR

PSNR is the most popular method for comparison of images in most of image processing and image coding fields. The popularity of PSNR stems from the

simplicity of computation. PSNR however does not give an accurate model of perceptual difference. The fact is verified in [27], [108] etc.

Since then, research in the scope of full reference perceptual quality assessment of images is divided into two approaches to solving the problem. The first approach is perceptually motivated, where the researchers attempt to model the HVS explicitly in some cases making assumptions on the nature of the images and attempt to predict quality. The second approach is structural similarity based, where change in structural similarity is used as a metric instead of a model based approach.

Perceptual Model based

One of the earliest attempts to form a quality metric is Mannos-Sakrison metric [71]. Here the authors tried to make use of Shannons rate distortion function and investigated a choice of distortion measure to simulate an optimum encoding. They accounted for luminance adaptation and CSF. This, however was limited to grayscale images.

A power spectrum based approach to image quality assessment is explored in [82]. The final quality metric is formed by transforming the image power spectrum by a modulation transfer function corresponding to the HVS. The metric works under the assumption that most natural images have a similar power spectral density.

The paper [27] describes two psychovisual experiments on with JPEG encoded images. In the initial experiment, images were compressed at different bit-rates and the users were asked to evaluate the quality of the distorted/compressed image with the original. 2AFC method was followed here with staircase increments. For the second experiment, the image was divided into blocks and paired ranking was used to compare image sub-blocks. The results of the user study was compared with the image quality metrics at that time namely, MSE, Logarithmic image processing metric(LIP), Distortion contrast (DCON),Mannos-Sakrison metric , mean intensity, spectrum slope etc. One of the major results of this study was that MSE was a poor predictor of user response. It showed high variability at the JND point. Distortion

contrast seems to be the best performer among the tested metrics. MANNOS that employs frequency weighting did not get significantly better results. Mean intensity proved to be a good predictor in supra-threshold experiment. The paper discourages the use of frequency domain methods as it require precise knowledge of viewing conditions.

Another area where exploration into HVS and image quality is in the field of compression of images. Some of the initial research in the field was done in [29], where major psychovisual phenomenon were explored in the context of image compression. Experiments were conducted here to study the effects of subthreshold error integration and masking at edges. A masking function was also derived here and was shown that it incorporated edge masking.

Another wavelet based approach to compression is [85]. The authors considers the limitations of HVS. The approach followed by the authors here was to initially study the response of HVS to sinusoidal gratings. Then they used HVS based wavelets that had a similar response as the one observed in studies of HVS. Followed by this appropriate bit allocation was done to achieve compression.

[114] is also a HVS model that can be used in image compression. It accounts for frequency selectivity, orientation sensitivity and contrast sensitivity of the image. The main idea used is the concept of local band-limited contrast (LBC). The definition of contrast used here is that of Peli [88] that considers ratios between frequency bands. Because of the use of LBC, the masking was found to be maximum at the location of an edge of the LBC transformed image. Furthermore, the threshold elevation was found to be approximately constant when LBC model was used. The final perceptual error value was found using

$$PEM = \left(\sum_{x,y} \left| \sum_{k,t} \Delta MLBC_{k,l}(x,y) \right|^{\alpha|\beta} \right)^{\gamma} \quad (2.1)$$

where α, β, γ are constants and $\Delta MLBC$ is the masked LBC value.

A sub-band coding approach is followed in **perceptuallytuned** It is a perceptually tuned image coding system. It decomposes the image into subbands and then adjusts the sensitivity of each sub-band on the basis of the local tex-

ture energy and intensity. The paper shows results proving that the masking threshold seems to follow an approximately linear increase with respect to difference in luminance (luminance edge) at a specific background illumination. It seems to increase. It also does a study on the effect of noise detection threshold at various illumination levels. The paper considers luminance masking by Webers law and masking by considering the luminance slopes of the neighborhood of each pixel. The JND is assumed to be independently contributed by both of these factors. The final JND is the weighted sum of JND of each of the sub-band.

Another approach to development of an image quality metric is to detects edge artifacts in images. This is studied in [48] . This is a full reference image quality metric that works by processing the errors between the original image and the reference image. The error image is modulated to account for masking effect due to high frequencies and background image brightness and orientation of edges. The study found that only a limited spatial background around the artifact affects the visibility. Another interesting result was that the presence of high frequency activity at a location can affect the visibility of the edges in the area. The authors tabulated the sensitivity of the edge against the various frequency. They also model the masking as a function of orientation and frequency.

A new approach to the problem was used in [117]. This is a different approach towards quality metric. It discusses a JND models based on self similarity. It uses the assumption that regular repeating patterns cannot conceal noise because of the predictability of image. The paper attempts to predict repetability of the structures by a structural similarity measure. It also accounts for luminance adaptation by using the threshold function defined in [A perceptually tuned subband image coder based on the measure of just-noticeable distortion profile]. The jnd threshold is calculated as a combination of spatial masking due to similarity and the luminance threshold weighted by the larger factor.

Quality metric can be formed by processing the difference between two images. A representative case in this case is [74]. This is a full reference

image quality metric that produces a score on the basis of differences in image, frequency and edges. It takes into account masking phenomenon, noise, other distortions like block noise etc. Taking all of these factors into account, five features are created, combined using PCA to produce a quality score. The metric claims to be able to produce scores with high correlations to human judgments.

One of the more popular models for image quality comparison is by Daly. This is an image quality model that give the probability that a difference will be noticed in various parts of the image. It takes into account luminance adaptation, contrast sensitivity and masking. Luminance adaptation is modeled locally. HVS frequency and orientation selectivity is modeled with cortex transform. HVS contrast sensitivity is modeled by a function that depends on radial spatial frequency, orientation, light adaptation and image size. The paper implements different models for masking in the final stages and has a psychometric model to determine the final probabilities of detection.

Structural similarity based

The second approach to image quality metric was proposed in [108]. This method is based on the fact that HVS is more sensitive to structural information in images rather than a threshold based approach. Hence the guiding principle here was instead of modeling every stage of the process, an overall effect (structural similarity in this case) would be evaluated.

The similarity in [108] is calculated by dividing images into patches and then comparing the luminance, contrast and structure of the two patches. The formula for overall SSIM is given by

$$c(X, Y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (2.2)$$

SSIM was further improved in [111] by introduction of multi scale SSIM and [97] that used SSIM in the wavelet domain. The paper claims to get very good results and is considered one of the leading publications in image quality metrics.

Other methods explored in the domain are [57] that tries to combine multiple methods of quality assessment. The logic behind this approach was that HVS might use multiple strategies to detect differences in images; for high quality, HVS looks for distortions in presence of image, and for low quality HVS looks at the image content in presence of distortions. For the first case, MSE error weighted by masking functions are used and for the latter, texture analysis approach is used where the images are passed through a gabor filter bank and a difference between coefficients are used as a similarity measure.

Based on earlier studies [108], it is believed that structural similarity plays a major role in perceptual quality assessment. Additionally, there are findings ([35] and [127]) demonstrating that low level features can influence visual perception and can model how humans assess image quality. Recent theories ([57] and [116]) also claim that the HVS in fact uses multiple strategies to detect differences in images. The first mechanism dominates when the distortions are not easily visible. Here the visual system seems to employ a detection based strategy, that most of the IQA algorithms, e.g., [21], apply to explicitly model how the HVS detects quality. The second mechanism becomes effective when the distortions are supra-threshold or visible. Here the quality is determined mostly by the perceptual ability to recognize the image content. Visual saliency is another research approach to study IQA. Zhang et al. claimed that differences in visual salience maps can be used as a predictor of visual quality [125].

Another study [67] claims that SSIM can be further simplified by just considering the gradient magnitude and choosing a better pooling strategy. The fundamental theory, however, is still based on the concept of structural similarity. An alternative method that has been successful is IFS [15] that tries to compute differences in terms of luminance and features extracted using Independent Component Analysis (ICA). ICA is also used to model the color mechanism of the HVS. Luminance distortion is calculated separately and combined with the ICA outcome to generate a final quality score.

Another recent work that tries to model IQA by learning algorithm is [45], which train a convolutional neural network (CNN) specifically to estimate

quality and distortion. In this case, the local error maps and the pooling are also determined by the training process of the CNN. However the method has the limitation of the model over-fitting on the trained dataset.

2.1.2 No reference Image Quality assessment

The problem solving approach followed in LDR NR-IQA uses some form of machine learning. Most of the methods start by creating a feature image by using different processing methods and then fit an arbitrary distribution on it. The parameters of this distribution are used as the feature vector of the distorted image (for example, DCT of the image would be utilized as a feature image and a Gaussian distribution would be fitted on the same, the features would be the mean and variance of the Gaussian). The features are then used as inputs to some learning system, which is used to generate a quality score.

One of the first machine learning based methods in LDR NR-IQA is BIQI [75]. It is a *two step process* where, from a set of features, an SVM predicts the type of distortion and another set of SVRs' predict the score for each kind of distortion. The final quality score is computed by

$$score = \sum_{i=1}^m p_i \cdot q_i \quad (2.3)$$

where p_i represents the probability of each distortion obtained from the SVM and q_i represents the quality score given by each of the SVR's. BIQI [75] used Daubechies 9/7 wavelet as feature image.

A variety of methods follow a similar approach and show very good performances on assessing LDR content without reference. Notable examples are BRISQUE [73], DIIVINE [76] and SSEQ [69]. These algorithms have more complex features and just need a single SVR to predict the final quality of an image.

BRISQUE [73] computes a mean subtracted contrast normalized (MSCN) image as feature by using $MSCN(i, j) = \frac{I(i, j) - \mu_{I, N, i, j}}{\sigma_{I, N, i, j} + 1}$, where $\mu_{I, i, j}$ and $\sigma_{I, i, j}$ represents the mean and variance computed over a local Gaussian window of size N around the point i, j . DIIVINE [76] uses divisive normalized steerable pyramid decomposition coefficients to create the feature image.

SSEQ [69] generates features using entropy as features. Here, a scale space decomposition is carried out to generate three scales of images, and then entropy is calculated for image blocks in the spatial and DCT domain. The entropies are then pooled by percentile pooling and the mean and variance of the spatial and frequency components are used as a feature vector.

An alternative approach that achieves state-of-the-art results in LDR NR-IQA is the Convolutional Neural Network (CNN) based approach used by Kang et al., kCNN [43]. This is a better implementation of the concepts introduced in [121]. The basic idea here was to learn discriminative features that can perform IQA rather than a handcrafted approach. [121] used dictionary learning to form discriminative filters. [43] improved this by redesigning it as a convolutional neural network (CNN). The CNN has four layers that act on MSCN image blocks of size 32×32 . The first layer is a convolutional layer with 50 filters (kernels), then a pooling layer that reduces the dimensionality of the data and finally two fully connected layers. The network is trained with the Mean Opinion Scores (MOS). The method has an additional advantage that it can produce an "error map" showing the visible errors on the distorted image.

An important observation here is that LDR NR-IQA algorithms rely on image distortions altering the statistical properties exhibited by 'natural' undistorted images. Hence, the problem considered in this domain of research is that of *quantifying the changes in natural image statistics*. Another key observation is that the natural statistics are not modeled explicitly but captured in the internal representation of the SVM or the CNN's used in the algorithms. Because of this, algorithms need an explicit training stage that adapts to the data before use. This training helps the learning component to learn what a "natural" feature is and how the noise changes this natural feature.

A recent work that tries to alleviate this problem is [129]. The research looks into assessing perceived quality of images corrupted by uniform and high frequency noise. The method uses a custom combination of feature whose weights that scales the error. This method however is not valid in our case as it cannot estimate compression errors.

2.2 High-level features in Image quality assessment

Research on image features has a long history, but the work on structural features was pioneered by Wang [110]. The concept was instead of modeling and evaluating every stage of the visual error recognition process, an overall effect of the error (structural similarity in this case) is evaluated. Subsequent work includes phase congruence [126], visual saliency [124] and a mixture of high- and low-level features depending on the apparent distortion [59], etc. All these methods focus on judging distortions and determining deterioration of quality with respect to a known reference. However, none of the methods comments on how the scene content may influence the image quality decision.

2.2.1 Visual saliency

Visual saliency is a capability of the HVS, where the attention of the human eyes is attracted to certain elements in the scene. It is a mechanism that helps to focus on more relevant or important information. Visual saliency is conventionally measured in user studies, where the users view an image freely in a given time. The eye movements are recorded with an eye tracker. The data captured by the eye tracker over time is expressed as a heatmap or a grayscale image called a saliency map [41].

The visual saliency process can be bottom-up or top-down.

- Bottom-up visual saliency means that the image saliency is determined by observing the low-level image features like edge orientation, color change, etc. Since the work of [41], a lot of approaches have been proposed to mathematically predict saliency maps for images. The latest modeling techniques include [9] and [66]. The models were found to agree with the earlier eye tracking findings.
- Top-down saliency suggests that an understanding of high-level content can influence the visual attention at a lower level. Recent research has looked into such top-down approach [119], as well as object dependent

visual saliency (salient object detection) [16]. These studies, however, have not analyzed the impact of top-down influence and visual saliency influence on image quality, which is the focus of our work.

Changes in Visual attention with task

The first study demonstrating changes to visual attention mechanisms was conducted by [120]. One of the major observations of the study was "Depending on the task in which a person is engaged, i.e., depending on the character of the information which he must obtain, the distribution of the points of fixation on an object will vary correspondingly, because different items of information are usually localized in different parts of an object". The authors arrived at this result by analyzing the eye movements of subjects asked different questions while viewing an image content.

The experiment shows the inadequacy of a free viewing eye tracker data towards evaluation of visual quality; the resultant fixation maps when the users are free viewing the image and when the viewers are performing quality assessment should be different.

2.2.2 Saliency and image quality

A related study looking into the association of visual saliency and image quality is [33]. The authors performed two experiments with eye-tracker. One with free looking task and another with quality assessment task. The tests were conducted on images in the LIVE dataset only ([99]).

They also found that in terms of improving Image Quality Metric (IQM), fixation points in free viewing were better, compared to fixation maps while doing IQA. The hypothesized reason for this was that IQA's already estimated the fixation on errors and adding extra weights would lead to over-estimation of the image quality scores. The paper did not delve further into the causes of these changes and concentrated on improving image quality metrics with free viewing fixation maps.

The authors found that the *visual attention maps obtained from the free viewing task and the visual quality evaluation task were different*. Further

investigation was conducted in their later study [32], but the authors focused on the free viewing task rather than the image quality assessment task. In contrast, we are interested in the quality assessment task. Furthermore, there is little research that looks into improving IQM with scene context. Most approaches focus on the use of image patch content or saliency, e.g., [128] and [96].

2.2.3 Image aesthetics and high-level features

There are many works that claim to assess image aesthetics using various image features. Most of them rely on adapting rules from photography to automated visual quality analysis ([6],[47]). Certain other works rely on using CNN's to model the system as a whole and predict a quality score. One of the latest work here is [46] which uses CNN-based method and image categories to learning techniques to predict the visual appeal of images. However, none of the techniques provide a good explanation of how features are significant on the basis of general image content.

2.3 HDR image quality

HDR images are images that use more than the conventional 8 bits to capture, store, transmit and display. Depending on the standard being used, this could be either 10bits or 12+ bits. HDR images, when viewed on a compatible HDR screen can show a wider range of colors and can display larger range of luminances compared to an LDR image, viewed on a conventional screen.

Image perception on HDR screens are different from that on an LDR screen. The changes in perceptual characteristics on HVS have been extensively studied by Aydin et. al. [7]. The study found that when viewing a distorted image on two different displays with different maximum luminances, the image shown on the brighter display was perceived as worse. The study further went on to prove that viewing content on LDR display and an HDR display has different perceptual effects due to differences in visual sensitivity at larger luminance ranges. Therefore, a direct statistical comparison of the errors with-

out considering perceptual changes in sensitivity with luminance may not give an accurate model for HDR data.

In terms of a full reference assessment of HDR image quality, an important work is HDR-VDP-2.2 [81] and HDR-VQM. HDR-VDP 2.2, models the early stages of HVS, based on psychophysical measurements. HDR-VQM uses sub-band decomposition and spatial and temporal error pooling. Both methods are full reference and outputs a local error visibility map as well.

In terms of no reference assessment of HDR image quality on HDR screens, we found no previous research. A search on the same topic often leads to another category of LDR No reference algorithms. These are tone mapped image quality.

2.3.1 Tone Mapped IQA

The research under this category focuses on evaluation of HDR images that have been tonemapped to LDR range, displayed on LDR screens. The field of research is relatively new with the first publication on FR-IQA for tone mapped images is [122]. The paper uses structural fidelity criteria akin to [111], however, alters the standard deviation based on a CSF. Along with this, a naturalness measure is also implemented by using methods similar to NR-IQA by fitting a Gaussian and beta probability density function on histograms of mean and standard deviation of the images. This was further advanced in [54], where the performance was improved with better error pooling and naturalness measure. Phase congruency was used as a feature in [78] for the same purpose. The phase congruency is added as an additional feature to compare the two images.

One of the only research that looks into NR-IQA for Tone Mapped HDR is [55]. The method uses MSCN as spatial domain features and gradient computations on different neighborhoods of every pixel. This is followed by Gaussian parameter extraction and SVR like the other techniques described here. The publication uses images tone mapped to LDR range by fusing LDR photographs of multiple exposures. The MOS scores are computed on an LDR display. The method is purely statistical and does not use perceptual

psychophysical modeling. This research is different from our work because our results are valid of HDR images displayed on compatible HDR display.

Tone-Mapped HDR is different from HDR. Experimental evidence of psychophysical differences in viewing HDR and LDR image content was provided by [2]. The study carried out a user study on how HDR and LDR image contents are perceived when displayed on HDR screens, with users rating naturalness, visual appeal, spaciousness, and visibility. The parameter of interest that is relevant to the current work is visibility. The study found statistically significant differences in how the users rated visibility in HDR and LDR images. Here visibility refers to the details in the image. Additionally, the authors found that aesthetically, the users preferred HDR images displayed on HDR screens compared to LDR images.

2.3.2 Algorithms for comparing performance of HDR NR-IQA

Because of lack of research in the area of a true HDR Noreference image quality assessment algorithms, we approach the problem of HDR FR-IQA through the use of the LDR-NRIQA on PU-encoded HDR data. These are based on the results of [26], that did a survey on the performance of various FR-IQA on HDR data. They found that PU encoded HDR data that is assessed by LDR-FRIQA algorithms performed well on HDR conditions. Hence we hypothesize (and confirmed) that these methods can provided a result that we can compare with.

2.4 Perceptual Quality of 3D textured meshes

Most of the works related to the perceptual quality of 3d models are limited to un-textured 3d meshes. Some of the earlier publications relied on simple properties of mesh like dihedral angle, Hausdorff distance and roughness [106], [18], [60] etc. A lot of these metrics are aimed at a specific task like simplification in [50], transmission of signals in **karni** watermarking in [20] and rendering in [24].

A different approach, where curvature was used in a manner similar to structural similarity [108] in image was [63]. This was later extended by [61] and [62]. These metrics make use of various statistical properties of a geometric mesh and model the perceptual quality. These measures are however limited to the structural properties of the mesh surface only and do not model the effect of quality of the texture on the data.

Most practical applications of 3D models however involves the use of textures. For certain applications like VR and medical data, the appearance of texture play a critical role and cannot be ignored. One of the first publications to work on perceptual quality of a textured 3D mesh was explored in the field of transmission was by [17]. A quality prediction model that takes geometry and texture into account was proposed here. ZIM predictors are used for evaluation of the texture quality. The predictor is a combination of the range required for color quantization Z (RGB color model), the texture intensity component I (HSI color model), and the degree of visual masking M , induced by the pattern complexity. The quality computation was done by taking a linear combination of both of these features, applying equal weights to both of the predictors.

[87] is one of the first attempts to take both texture and geometry into account while estimating the quality. The method attempts to fit a curve that satisfies the observed quality levels under different texture and mesh resolutions. The model was a global approach and is very fast as it does not require complex calculations. It finds use in quality control while transmitting in a limited bandwidth. This model is however restricted to effect of texture resolutions. The effect of compression is an area where more research needs to be done.

2.5 Datasets used for evaluation of results

To validate our results on different imaging modalities and distortion types, we make use of a large number of datasets available in the public domain. We use a standard dataset containing a set of images, its degraded versions and

the perceptual quality scores (mean opinion scores) determined by user studies on those set of images. We generate a quality score on the images using our algorithm. Then we compare the scores by the algorithm with that of the user quality scores in the dataset. The comparison is performed using Spearman rank order correlation coefficient (SRCC), Kendall rank-order correlation coefficient (KRCC) and Pearson linear correlation coefficient (PLCC) and root mean square error (RMSE). A good performance is indicated by high values of SRCC, PLCC and KRCC and low value for RMSE.

2.5.1 Low dynamic range datasets

The datasets I use for evaluating the performance of LDR image quality assessment algorithms developed in this thesis are CSIQ, LIVE, TID2008 and TID2013.

The LIVE dataset (Fig. 2.1) has 29 reference images, distorted using five distortion types. JPEG2000, JPEG, white noise in the RGB components, Gaussian blur, and transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model. CSIQ dataset has a total of 866 distorted images are available. For the quality assessment task, observers were asked to provide their opinion of quality using “Bad”, “Poor”, “Fair”, “Good” and “Excellent”. About 20-29 human observers rated each distorted image by comparing with the reference images.

The TID2013 contains 25 reference images and 3000 distorted images (25 reference images x 24 types of distortions x 5 levels of distortions). Reference images are obtained by cropping from the Kodak Lossless True Color Image Suite. The TID2008 dataset was the first version of TID2013 and is included for compatibility of results with older publications.

2.5.2 High Dynamic Range datasets

For a comprehensive data set of HDR image, we selected 5 separate data sets. The authors of the respective publications performed subjective evaluations using different displays with varying maximum intensities and viewing distances providing a good testing platform for the algorithms. [79], consisting



Figure 2.1: Example images in the LIVE dataset.

Database	No. of Images	Distortion Type	Resolution
CSIQ [58]	866	6 Distortions	512 x 512
IVC[65]	185	4 Distortions	512x512
LIVE[37]	779	5 Distortions	Mixed
TID2008[91]	1700	17 Distortions	512 x 384
TID2013[89]	3000	24 Distortions	512 x 384

Table 2.1: LDR database statistics.

of JPEG compressed HDR images, [80] consisting of JPEG2000 compressed HDR images, [52] with JPEGXT compression and [105], [26] containing images distorted by JPEG, JPEG2000 and JPEGXT compression schemes. The statistics of the data-sets are given in Table 3.1.

Chapter 3

Quality assessment of High Dynamic Range Images

3.1 Introduction

High dynamic range (HDR) imaging can capture a much larger range of luminance compared to conventional images. This larger range of luminance is achieved by increasing the number of bits used to store the image pixels. Typically HDR image content is stored using 10-12 bit floating point values per pixel as opposed to a conventional image file that uses 8 bits per integer for a pixel. HDR imaging technology incorporates techniques of capturing, processing and displaying HDR content. This technology is becoming mainstream in the consumer market and is reaching larger sections of people as a result of investments in the technology made by the camera and television manufacturers.

In order to reproduce native HDR content, an HDR display is required. HDR displays have the capability to display a very high range of pixel intensities. When viewing an HDR image through an HDR display, a viewer will perceive an increased range of colors and image details compared to a conventional screen displaying the same content. This results in a better quality of experience for the user. In this paper, we consider this case of HDR images being displayed on an HDR compatible display.

The process of quantifying the visual quality perceived by a human being is called image quality assessment or IQA. IQA can be mainly categorized into

Full-Reference (FR) and No-Reference (NR). In FR-IQA, the quality of the image is evaluated on the assumption that an undistorted version of the same image is available. In NR-IQA, the quality of the content is evaluated on the basis of the distorted image only.

IQA of the commonly seen image distortions on HDR images is a problem of great academic and industrial interest. Since the target audience for the HDR content is a human being, the easiest method for IQA is through a subjective test. However, subjective tests are often tedious and time-consuming. Even with massive crowdsourcing projects (systems like mturk), HDR IQA is difficult in view of expenses involved in acquiring systems capable of displaying the HDR content. Hence there is a need for high performance automated systems that are capable of IQA.

We propose the first model for NR-IQA that is capable of predicting the perceived image quality and localizing the distortions present in an HDR image. We use a convolutional neural network (CNN) based architecture to achieve this. The scope of this work is limited to commonly seen low-level image distortions like artifacts caused by image compression. We do not consider changes in image quality due to high-level changes, e.g., artistic intent, where complex aesthetics considerations should be taken into account.

The contributions of this work are as follows:

1. We propose the first HDR NR-IQA method based on a convolutional neural net architecture capable of separation of error and perceptual effects present in a distorted image. The proposed method outperforms other NR-IQA methods and is competitive with state-of-the-art HDR FR-IQA algorithms .
2. We provide a method for the accurate estimation of error in a distorted image without a reference.
3. We predict the perceptual masking factors error without an explicit psychophysical model.

3.2 Motivation

The research on NR-IQA discussed above is focused on LDR images displayed on a conventional screen with a maximum luminance of about 300 cd/m^2 . HDR displays have luminances up to 4000 cd/m^2 and more. The response of the HVS under this increased luminance range is radically different from the responses in LDR luminance range. Note that Tone Mapped HDR also comes under the category of LDR luminance range because of the LDR display used in the user evaluations.

Change in sensitivity in HVS response with luminance levels lead to loss of performance under HDR conditions. Therefore, a pure statistical modelling alone is not enough for an NR-IQA on HDR data. We confirm this fact with our experiments in section VI.B. We do not know of any literature in the field of NR-IQA that anticipates and adapts to these changes.

3.2.1 Perceptual factors affecting HDR data

The changes in perceptual characteristics on HVS have been extensively studied by Aydin et. al. [7]. The study found that, when viewing a distorted image on two different displays with different maximum luminance, the image shown on the brighter display was perceived as worse. The study further went on to prove that viewing content on LDR display and an HDR display has different perceptual effects due to differences in visual sensitivity at larger luminance ranges. Therefore, a direct statistical comparison of the errors without considering perceptual changes in sensitivity with luminance may not give an accurate model for HDR data.

Hence, we argue that, in addition to statistical modeling, HDR NR-IQA requires considering the psychophysical phenomena that determine the perception of distortion in HDR conditions. Even though such a perceptual approach is not very popular in NR-IQA, it is very popular in FR-IQA models. An important work in this category of algorithms on HDR data is HDR-VDP-2.2 [81] and HDR-VQM. HDR-VDP 2.2, models the early stages of HVS, based on psychophysical measurements. HDR-VQM uses subband decomposition

and spatial and temporal error pooling. Both methods are full reference and output a local error visibility map as well.

3.3 Proposed method

Designing a perceptual model is a challenging task because of the lack of understanding of how the existing psychophysical measurements with sinusoidal gratings generalize to a complex image. The traditional approach to this problem considers various visual features like contrast, frequency and background luminance etc into account. Then, a handcrafted function is used to combine all these features into a quality score. This function is derived from fitting functions to the results of various psychophysical experiments which test the limits of our perception. Generalization of these results to a real-world image is, however, a problem of interest and yet to be solved.

We approach this problem of designing a perceptual HDR NR-IQA by dividing the visual quality perception process into sub-components and modeling them individually. Conceptually, the visual quality perception can be represented as an interaction of two functional units. The first unit takes distorted image as input and detects error, and the second unit performs a perceptual scaling of this error to compute a quality score.

We first model the error detection unit. Then, using this result, the distorted image data and real-world MOS scores from IQA databases, we perform an optimization that derives the perceptual component and the quality score simultaneously. Differently from existing perceptual IQA methods such as HDR-VDP, we find the perceptual component in our network directly from data.

3.3.1 Design

To model the above idea, we design a convolutional network architecture that acts on image blocks of HDR linear luminance values. We consider a block size of 32x32 pixels. We consider a block size of 32x32 pixels corresponding to the pixels in one visual degree under standard HDR test conditions. This

value was also suggested by [44] in their CNN framework. The pixels are then processed by a CNN model whose architecture is depicted in Fig 3.1. As part of this model, we design three systems. The first system, E-net, works on estimation of *Error* contribution $\delta(i, j)$ of a given image block centered at i, j . The second system, P-net, computes the combined perceptual effects of the block; we refer to this as the *Perceptual Resistance* of a block $T(i, j)$. The output of these two systems are then combined using the third system, a *mixing function*, to produce the local quality (we use Difference in Mean Opinion scores or DMOS computed as $100 - MOS$ in our datasets) of the image block. The block scores are then combined to form an overall quality of the image.

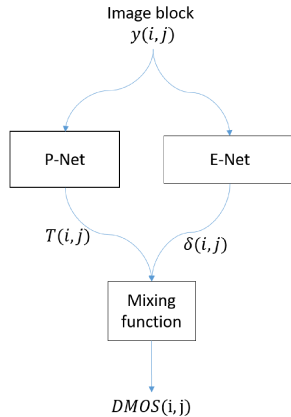


Figure 3.1: Proposed strategy for IQA. E-net detects the error, P-net detects the perceptual response, Mixing function combines the results to produce a DMOS.

3.3.2 Error Estimation

The Error $\delta(i, j)$ represents a number that quantifies the corrupting statistics present in the image block. For an image block centered at (i, j) , we define the error of the block as,

$$\delta(i, j) = \text{mean}(|Y_R(i, j) - Y_D(i, j)|) \quad (3.1)$$

where Y_R and Y_D represents the the undistorted and distorted linear HDR luminance values of image block centered at the point (i, j) . Note that this

is not a Full Reference Algorithm. The undistorted version here is only used during training; any HDR image and its distorted version can be used here. The objective here is to give representative examples of characteristic patterns produced by image distortion like blocky artifacts, blur, jagged edges etc. Conceptually, E-net would learn and quantify these patterns.

The choice of L1 error best represents the overall impact of the error of the block, giving equal importance to extremely high and low values of error (as opposed to biased values obtained the L2 norm is used).

To estimate this error, we use our own CNN, E-net, with the training data being linear luminance values of the image block centered at any (i, j) and the training target being to reproduce the average error in the block computed by equation 3.1.

3.3.3 Perceptual Resistance

For each image block centered at (i, j) , we compute the *Perceptual Resistance* $T(i, j)$. This represents how difficult it is for a viewer to perceive the error of the block $\delta(i, j)$. A high value for T implies that a subject is less likely to see the error of the block, hence the quality of the block will be less affected. A low value implies that the image block will be perceptually degraded by error. The purpose of this term in our system is to perceptually scale the error $\delta(i, j)$ in order to produce a quality score through a mixing function.

Perceptual Resistance $T(i, j)$, represents the combination of all the perceptual effects exhibited by the image luminance block centered at (i, j) . Though functionally similar to the pixel-wise Just Noticeable error thresholds (JND) in conventional IQA systems like [129], [21] and [81], we define Perceptual Resistance as a new term. This was done to differentiate the fact that the results given by our model are local quality scores (DMOS) as opposed to a local probability of error detection.

An important detail worth mentioning here is that JND is traditionally determined by mathematical modelling of the response of the HVS to the region surrounding any location of interest. Various visual features like contrast, frequency and background luminance are considered and custom functions are

used to combine these factors. These functions, in turn, are derived from various psychophysical experiments. The experiments check how HVS responds to simple stimuli under various conditions. It cannot be assumed that these methods can be generalized to a real world image with complex frequencies, luminance and contrast levels.

Our solution to this problem is instead driven by data. We use a convolutional network based architecture, P-net, to derive the Perceptual Resistance of the block. The *CNN analyzes the image content and computes the features required to do this task from real world image data* provided to it by numerical optimization. Thus, the Perceptual Resistance is computed by a neural network, whose behavior is guided by a potentially large number of 'functions' (that are represented as neural network weights) that best explain the observed data.

3.3.4 Mixing Function

To combine the error and Perceptual Resistance, we use a *mixing function*, represented as $f(\delta, T)$. This is an important part of the formulation since it determines the behavior of P-net. The result of the mixing function would be optimized by the training process to match quality score; hence depending on this function, the output of P-net would change. For example, if we choose $DMOS = \frac{\delta}{T}$ and the CNN training optimization converges successfully (error values does not decrease with training in training and validation sets), P-net would generate a T that acts as an error detection threshold to the value produced by E-net.

While it can be argued that the mixing function can be chosen by another CNN, this process would involve more weights and difficulties in optimization resulting in the overall model not converging to a good solution. We confirmed this fact experimentally by using a DBN (Deep Belief Network) in place of the mixing function. Even if the system did converge with a DBN, this function would be a 'black box' with no intuitive interpretations.

We design the structure of the mixing function to express error in multiples

of Perceptual Resistance:

$$DMOS = f(\delta, T) = G\left(\frac{\delta}{T}\right) \quad (3.2)$$

where $G()$ is some arbitrary function. To observe how $\frac{\delta}{T}$ is indicative of the quality, consider a value of $\frac{\delta}{T} < 1$; from the low value of the ratio, we can infer that the error is very small compared to the Perceptual Resistance of the block, implying that quality of the block is very high, hence a lower DMOS. Conversely, $\frac{\delta}{T} > 1$ would imply a higher DMOS. Such formulations are frequently used in perceptual image quality literature like [129], [7] and [81] etc. The common step in all these publications is expressing error in JND units ($\frac{error}{JND}$). The objective of this step is to convert the error into a more perceptually relevant value.

Translation of $\frac{\delta}{T}$ to quality scores is achieved by the function $G()$. From the example above, we see that DMOS has to increase with this ratio, implying that $G()$ will have to be monotonically increasing with $\frac{\delta}{T}$. Other than this constraint, the choice of $G()$ can be arbitrary; any function that is monotonically increasing would be sufficient as long as the optimization converges.

However, choosing a $G()$ that is too complex can also lead to optimization problems because of unstable points along the function or low values for gradients, leading to slow or zero learning. We do not go into the mathematics of CNN convergence and optimization functions as it is beyond the scope of this work.

Based on the above considerations, hereafter we use $G(x) = 1 - \exp(-|kx|)$ for obtaining DMOS. Here k is an adjustable parameter to control the shape of the function. Hence the DMOS of an image block centered at (i, j) is defined as

$$DMOS(i, j) = 1 - \exp\left(-\left|\frac{k * \delta(i, j)}{T(i, j)}\right|\right) \quad (3.3)$$

This choice is inspired by the error model proposed in [129]. We modified this equation by adding a scaling factor k to control the shape of the function.

We found that similar results, with slight variations in performance, can be obtained by using other functions that are monotonically increasing like a

$G(x) = \tanh(kx)$ function or a logistic function $G(x) = \frac{1}{1+\exp(-k(x-x_0))}$; where k , x_0 are parameters that control the shape of the function. In practice, the value $T(i, j)$ from P-net would change according to the function used.

As seen from the plot of Eq 3.3 in Fig 3.2, the function represents a wide range of values and rate of increase in DMOS, depending on the value of $\delta(i, j)$, $T(i, j)$ and k . In our method, the value of k is determined as part of the optimization.

The expected behavior is captured by this formulation. For example, for the Perceptual Resistance $T(i, j)$ that is large, the output of the function would remain small even if there is a large error $\delta(i, j)$, e.g. $T = 200$ in Fig 3.2 A . This scenario would be seen in cases of high masking because of luminance or complex patterns.

Conversely, if the Perceptual Resistance is low, the predicted DMOS would shoot up very quickly even for the smallest of error. Practically, this would be seen in areas that are smooth with very low masking effects.

k serves to control how slowly the DMOS can increase with $\frac{\delta}{T}$ (see Fig 3.2 B). This parameter remains constant after training and does not change with image content.

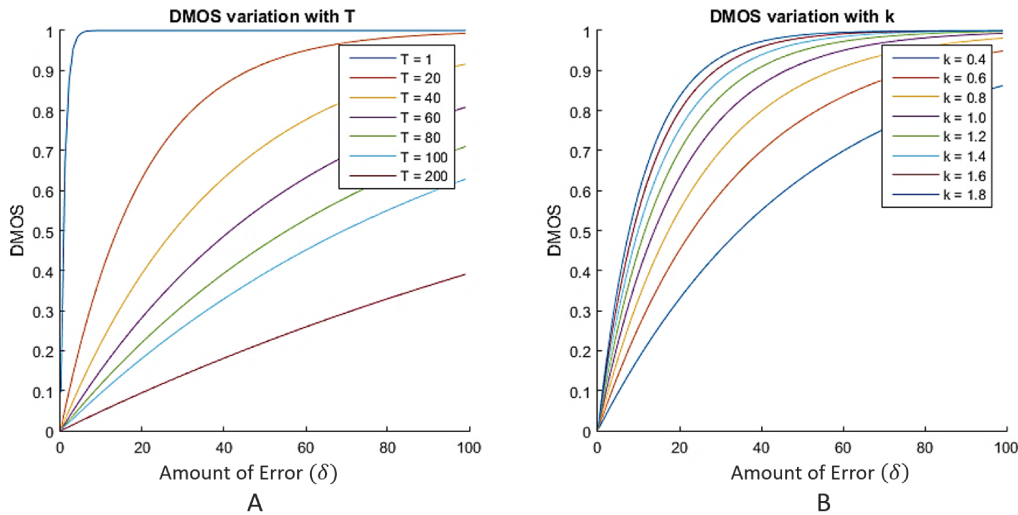


Figure 3.2: Behavior of the mixing function. (A) Varying T with k fixed at 1. (B) Varying k with T fixed at 20.

The block-wise DMOS scores obtained from equation 3.3 is converted to

global scores by simple averaging. A weighted scheme is not required here since the perceptual scaling of errors based on content is already handled by equation 3.3 (the term T computed by the CNN changes with image content and handles content-dependent scaling).

3.3.5 Training

The biggest drawback of a CNN based system is the large amount of training data required to compute the weights of the neural network.

E-Net estimates error; the training data for this task can be easily obtained. For each block of the image, the target value would be the mean error in that block. This, in turn, is the difference between the distorted and reference images (Equation 2).

For training P-Net, the ideal training data would be a number that combines all the perceptual effects of the HVS acting on the image block. We do not have this data, however, we can obtain the value for the final block quality after the mixing function. We use this for training.

In training P-net, we make a strong assumption that the block quality is equal to the global image quality score. Even though this assumption is incorrect most of the time, it was shown in [45] that, with a starting assumption that the global quality of the distorted image is the same as that of the local quality, the training process of the CNN isolates the local quality. A study of the filters generated by the CNN in the study revealed that it detected spatial patterns of error in the feature image provided. Under the assumption that the block quality is equal to the global image quality, multiple quality scores might be associated with the same pattern of error. However, when trained over millions of blocks with a cost function that imposes sparsity constraint ([45] used the L1 distance between the predicted quality and the actual quality), the correct local quality is the only value that minimizes the total error. In other words, the lowest cost of the CNN cost function is obtained when the CNN generated the true local errors of the image, regardless of the labels it started out with.

We now define our *two-stage training process*. In stage 1, E-net is trained

with image blocks as input and the corresponding mean error of the image block as a target, hence E-Net learns the patterns in the data corresponding to error.

Then, in stage 2, all the training weights of E-net are frozen by dropping the learning rate of this section to zero. The whole network is then trained with image block as input and global image quality ($DMOS_{global}$) score as a target. We use $|DMOS(i, j) - DMOS_{global}|$. The choice of L1 norm was inspired by the results of [36], who found that the regularizing properties of L1 norm helped achieve high performance in LDR NR-IQA. The process is illustrated in Fig 3.3.

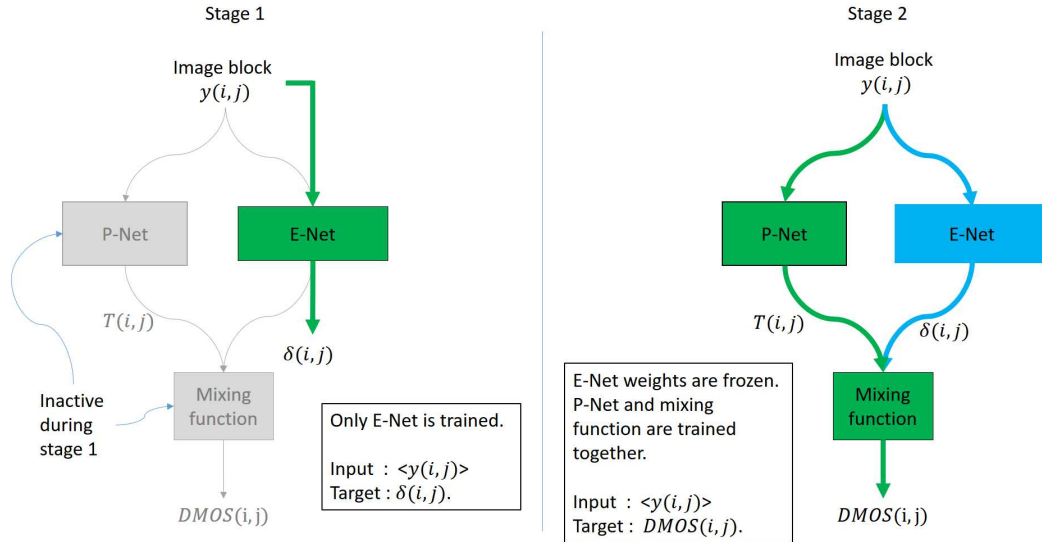


Figure 3.3: Two-stage training Process.

Note that there is no explicit training of P-net. The results of E-net, image block and the ground truth DMOS scores from image dataset are used for a global optimization. The cost function used for this optimization is the L1 norm between generated DMOS and the ground truth DMOS. It is this sequential training that forces the P-net to extract a set of perceptual features from the image block and derives a single Perceptual Resistance value from it.

3.3.6 Further considerations

By separating the whole process into sub-components as opposed to a single system we force the individual sub-components to model a simpler process. E-net handles modeling error statistics *only* and deals with detection of non-natural statistics in an image. P-net models perceptual processes *only* and produces an approximation of perceptual response and nonlinearity from image data. Mixing function deals with the conversion of the results of E and P-net to DMOS.

Thus, each system has to model only one physical process implying that the data to each sub-system have similar underlying distributions. The CNN's eliminate the need for rigidly handcrafting the behavior of individual sub-systems.

This two-model mechanism has two major advantages. First, a separation of perceptual components from the physical error gives us more intuitive results that can be used in applied fields of IQA like image and video compression. With adequate calibration, Perceptual Resistance values can be used to optimize compression or transmission. Secondly, it simplifies the learning process leading to *improved results* and *better generalization of those results* to real world cases.

3.4 Architecture

The proposed network architecture for the error estimation (E-Net) has 5 layers. E-Net is required to do a blind error estimation. Hence we choose a typical CNN architecture consisting of 5 layers. The layers are convolutional with 64 filters of dimension 7×7 , 128 filters of 5×5 , 256 filters of 3×3 and 512 1×1 filters. Spatial pooling of 2×2 was used after each filtering stage. The final layer consists of one node corresponding to the output. Spatial dropout layers [103] are added to prevent over-fitting of the data. The network structure is shown in Fig 3.4.

P-Net is required to estimate the Perceptual Resistance values of the block. Here, we define a custom CNN layer, Augmented Input Layer. In this layer, in

addition to the original luminance values of the block, we compute the mean, variance and MSCN images. The variance image is computed by replacing every pixel (i, j) with variance computed over a local Gaussian window of size N around the point i, j . For MSCN image, we use the equation proposed in [73], $MSCN(y_N(i, j)) = \frac{y_N(i, j) - \mu_{y_N(i, j)}}{\sigma_{y_N(i, j)} + 0.01}$. $\mu_{y_N(i, j)}$ is computed by replacing every pixel (i, j) with the mean computed over a local Gaussian window of size N around the point i, j . We use a smaller value for the stabilizing constant to prevent the stabilizing constant from influencing the MSCN values. Since a neural network training requires that the input value is in a similar range, we scale the input, variance map and the MSCN map with trainable weights whose values are determined as part of the overall optimization process. Hence the output of the augmented layer will be $[\langle W_1 * y_N(i, j) \rangle, \langle W_2 * \mu_{y_N(i, j)} \rangle, \langle W_3 * \sigma_{y_N(i, j)} \rangle, \langle W_4 * MSCN(y_N(i, j)) \rangle]$.

The succeeding layers consist of convolutional layers with 64 filters of dimension 3X3 and 128 filters of 3x3. This is followed by 2 densely connected layers with 100 nodes each. The final layer has one node corresponding to the output. The network structure is shown in Fig 3.5.

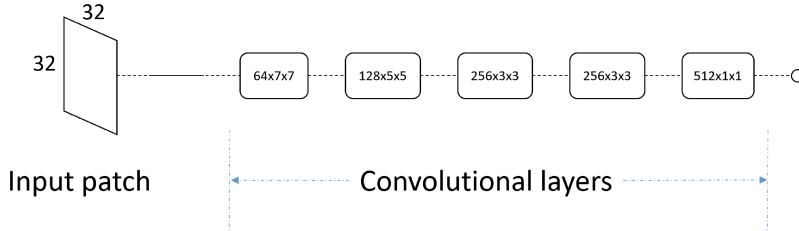


Figure 3.4: Network structure for Error estimation.

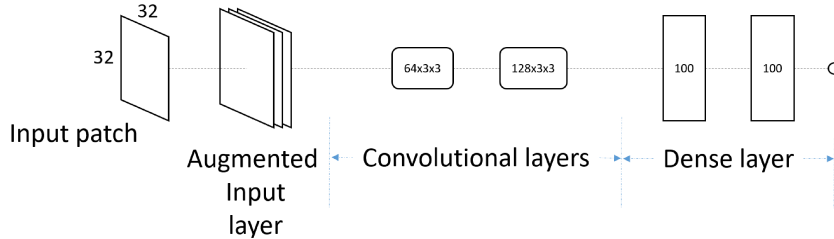


Figure 3.5: Network structure for Perceptual Resistance value.

The results of the two networks are combined by the mixing function whose behavior can be modelled by equation 3.3. Here the parameter k is tuned as part of the training process.

3.5 Results

The proposed algorithm was implemented on a computer with an Intel core i7 processor, 16GB RAM, and a Nvidia GTX660 graphics processor. The language used was Python with keras on theano backend, imageio and open CV as supporting libraries. We used the Adam optimizer ([51]) to optimize the weights of the CNN. The parameter values of Adam was learning rate=0.001, $\beta_1=0.9$, $\beta_2=0.999$, $\epsilon=1e-08$ and decay=0.0. The batch size used was 200. The training was done for 10 epochs.

3.5.1 Dataset

For a comprehensive data set of HDR image, we selected 5 separate data sets. The authors of the respective publications performed subjective evaluations using different displays with varying maximum intensities and viewing distances providing a good testing platform for the algorithms. [79], consisting of JPEG compressed HDR images, [80] consisting of JPEG2000 compressed HDR images, [52] with JPEGXT compression and [105], [26] containing images distorted by JPEG, JPEG2000 and JPEGXT compression schemes. The statistics of the data-sets are described in table 3.1 .

The dataset provides MOS values for the images. Since our models expects differences in quality, we use a value $\frac{(120-MOS)}{120}$ for training the complete system training.

3.5.2 Reference IQA schemes

Due to lack of research into NR-IQA on HDR images, we test our algorithms on the major LDR NR-IQA (BRISQUE [73], SSEQ [69], BIQI [75], DIIVINE [76], and kCNN [43]) with and without various pre-processing operators. The results were obtained after retraining the algorithms on the respective processed

Dataset Number	Number of Reference Images	Number of Distorted Images	Distortion type	Maximum Luminance (Cd/m ²)
#1 [79]	27	140	JPEG	1000
#2 [80]	29	210	JPEG 2000	1000
#3 [52]	24	240	JPEG-Xt	4000
#4 [105]	15	50	JPEG JPEG2000 JPEGXT	4000
#5 [26]	15	50	JPEG JPEG2000 JPEGXT	4000

Table 3.1: Database statistics.

HDR data. The pre-processing operators we choose are PU encoding and various tone mapping including Reinhard 2002 and 2005 ([94], [93]), Drago [23] and Mantiuk [72]. PU encoding has been shown to perform well in a similar context in the case of HDR FR-IQA [105].

The features were extracted using the implementation provided by the authors. In the case of SSEQ [69], we normalized the images by the maximum intensity under the respective schemes (4000 for linear HDR and 455 for PU encoded data). For training the SVM, methods suggested by the authors were used (SVR with RBF kernel). A grid search on the cost and the kernel parameter of the SVM was conducted for a range of 10^{-15} to 10^{15} before training. [36] was re-implemented using python. For testing the algorithms, we used the same procedure as in [36], i.e. 100 iterations of training and testing with median scores of test cases reported. Note that the results can vary slightly since the weight initialization of CNN is random. Here, training and testing sets are formed ensuring that *there is no overlap in image content*, i.e. the reference images for the distorted images in training and testing sets were not the same.

3.5.3 Learning performance

To analyze the ability of the model to learn from the data, we perform the following experiment. We randomly sample one case of the training data and testing data from above (570 training images, 120 testing). We then train the model with 32 images (1000 patches per image). We repeat the same till every image in the training set has been visited at least once (1 epoch). The plot of the resulting training error vs batch number is shown in fig 3.6.

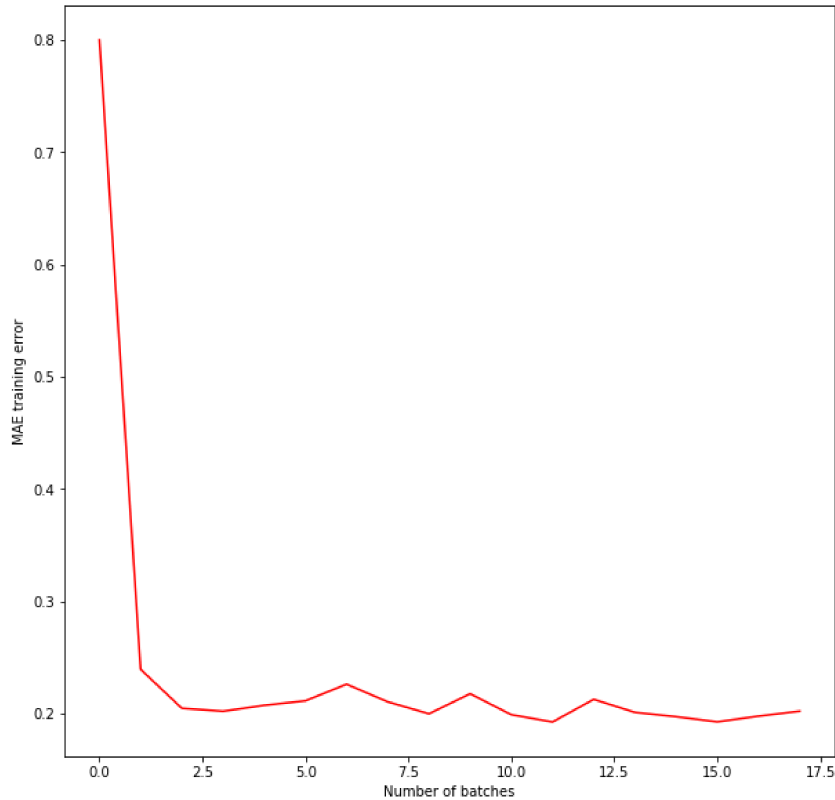


Figure 3.6: Training error vs batch number for every image in randomly sampled training set (570 images).

For the testing performance, we show results of training stage 2, i.e. we start after E-net is fully trained and we have to train the P-net. Performance with Mean Absolute Error on the test set would be meaningless as we intend to model the perceptual similarity. Hence, we show SRCC performance on the test set after training with each batch of 32. The results are shown in figure 3.7

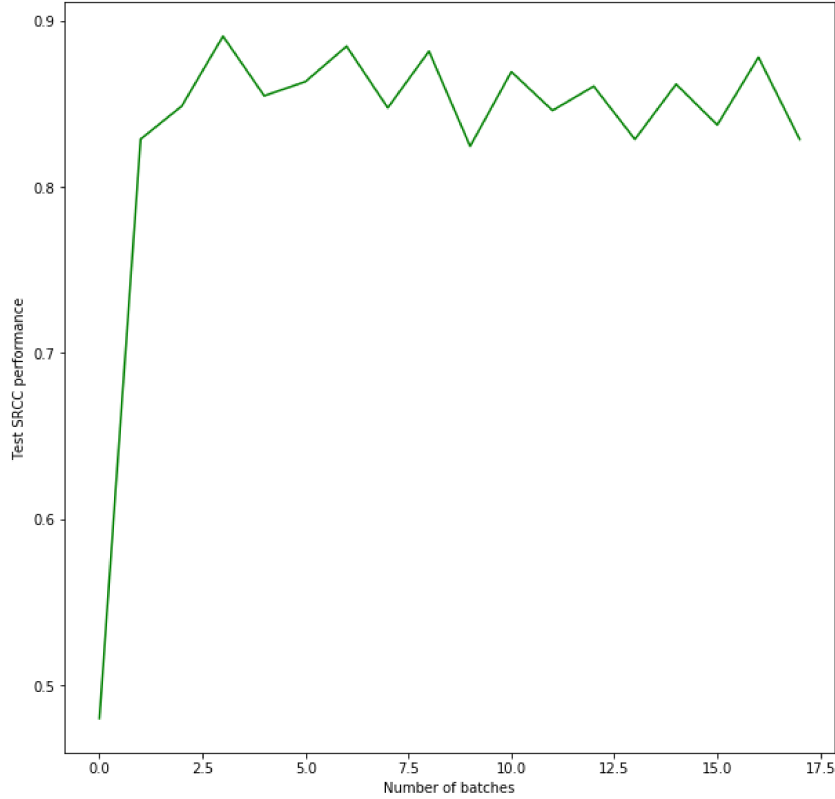


Figure 3.7: SRCC performance after training with increasing number of samples.

We see the results as we expect. The training error goes down and stabilizes to a low value and the testing performance (SRCC) goes up and then oscillates around the value of 0.86. Note here that this is not the final performance score since we only plot the performance on the first epoch. Training for more epochs would continue this trend and generate larger SRCC and lower training errors.

3.5.4 Performance comparisons

Comparison of performance was based on Spearman rank order correlation coefficient (SRCC), Kendall rank-order correlation coefficient (KRCC) and Pearson linear correlation coefficient (PLCC) and root mean square error (RMSE). A good NR-IQA is characterized by a higher value for SRCC, KRCC, and PLCC and a lower value for RMSE.

For the computation of SRCC, KRC and PLCC, the outputs of the proposed system was used directly. The RMSE for the proposed system was computed

using the formula:

$$\sqrt{\text{mean}\left(MOS_{MAX} * DMOS_{pred}(i) - (MOS_{MAX} - MOS(i))\right)^2} \quad (3.4)$$

for all image i in the dataset.

The results in terms of SRCC, KLCC, and PLCC are shown in Tables 3.2.

Considering the NR-IQA originally designed for LDR content, we see an acceptable performance after retraining with SRCC around 0.7 for many of the algorithms; as hypothesized, the learning component of conventional NR-IQA is adapting and capturing the statistics of the error types even in HDR conditions. Best performances were obtained by using BRISQUE [73] and kCNN [45]. The high performance of BRISQUE and kCNN can be attributed to the features they use i.e. the MSCN coefficients. It is likely that the normalization by variance acts to cancel the effects of the increased dynamic range and yields a similar distortion pattern as LDR images. Practically, kCNN is more useful because it produces an error map in addition to the quality score. The error map shows an approximate error that is seen by the observer on the noisy image.

Furthermore, we observe a clear performance improvement in LDR-NR-IQA algorithms if the data is pre-processed and the dynamic range of the data is reduced to LDR levels. PU encoding improves the performance in most of the cases. The best performance among LDR-NR-IQA was obtained in the case using PU encoding in conjunction with kCNN.

The performance of the proposed system is significantly better than the other algorithms in all cases even after compensation with PU encoding.

Statistical Analysis

To further validate our approach, we perform statistical analysis on algorithms that show performance similar to ours (Bolded in table 3.2).

We follow the approach used in [73]; here, we use different randomly selected training and test sets and record the SRCC performance for each of the training set used. For our tests, we re-run the training of the four algorithms

Feature	Processing	SRCC	KRCC	PLCC	RMSE
BRISQUE	Lin	0.7274	0.5430	0.7231	18.1797
	PU	0.8047	0.6116	0.7825	17.3576
	TMO - Drago	0.7374	0.5415	0.7203	19.1261
	TMO - Reinhard 02	0.7782	0.5853	0.7699	18.1523
	TMO - Reinhard 05	0.6903	0.5061	0.6643	20.3307
	TMO - Mantiuk	0.6172	0.4559	0.6148	22.1868
SSEQ	Lin	0.6022	0.4378	0.6008	23.3017
	PU	0.7342	0.5451	0.7175	19.4117
	TMO - Drago	0.6853	0.5011	0.6954	20.8766
	TMO - Reinhard 02	0.6866	0.5183	0.6688	21.0673
	TMO - Reinhard 05	0.6568	0.4845	0.6467	20.5737
	TMO - Mantiuk	0.4185	0.2926	0.4651	25.7570
BIQI	Lin	0.1817	0.1391	0.1466	38.7513
	PU	0.3387	0.2406	0.3445	30.5220
	TMO - Drago	0.2803	0.1923	0.2960	41.0579
	TMO - Reinhard 02	0.3756	0.2778	0.3766	33.2005
	TMO - Reinhard 05	0.3097	0.2213	0.2874	27.7294
	TMO - Mantiuk	0.2822	0.1957	0.2408	39.0999
DIIVINE	Lin	0.6677	0.4853	0.6759	21.8020
	PU	0.7156	0.5290	0.7193	18.7586
	TMO - Drago	0.7418	0.5562	0.7400	18.9959
	TMO - Reinhard 02	0.7149	0.5266	0.7024	20.7177
	TMO - Reinhard 05	0.7900	0.5932	0.7809	17.2134
	TMO - Mantiuk	0.4946	0.3549	0.4936	27.4918
kCNN	Lin	0.8363	0.6530	0.8134	19.1753
	PU	0.8638	0.6852	0.8497	16.8937
	TMO - Drago	0.7700	0.5853	0.7485	18.2759
	TMO - Mantiuk	0.8075	0.6188	0.8053	17.7948
	TMO - Reinhard 02	0.8613	0.6668	0.8179	17.7157
	TMO - Reinhard 05	0.6438	0.4631	0.6074	22.3484
Proposed	Lin	0.8920	0.7184	0.8860	14.1464

Table 3.2: Overall Performance comparison.

20 times. For a fair comparison, each of the algorithm was trained for an equal amount of time.

The results were then compared for significant differences using the one sided t-test. The null hypothesis was that the mean correlation of the algorithms are same, the alternative was that the mean correlation is different. We reject the null hypothesis if our $p < 0.05$. We can indicate if the difference is greater or smaller by using the mean correlation of each method. We denote it similar to [73] as (1) if algorithm in the row is better than the one in column, (0) if no difference in mean and (-1) if mean is lesser. Results are recorded in table 3.3.

We can see that the proposed method does indeed produce a statistically

	Proposed	kCNN	PU-kCNN	Reinhard02-kCNN
Proposed	0	1	1	1
kCNN	-1	0	-1	-1
PU-kCNN	-1	1	0	0
Reinhard02-kCNN	-1	1	0	0

Table 3.3: Statistical significance of difference in performance of algorithms (Bolted in table 3.2).

significant improvement in performance.

3.5.5 Generalization capability

One of the common complaints against an NR-IQA system is that the performance cannot be generalized in situations with a different conditions and contents. To test this, we train the algorithms using data sets #1,#2 and #3 and test it on #4 and 5. This represents a *real-world test* scenario where the experimental conditions are different as that of the training data. In addition, this testing method also allows us to perform a head-to-head real world test against the performance of full reference image quality assessment algorithms. From a machine learning point of view, this is acceptable, since we have sufficient number of examples of each type of distortion in data sets #1,#2 and #3 and a combination of all of the distortions in data set #4 and 5. The test set contains DMOS scores uniformly distributed in the range [20,90].

Since the CNN are initialized with a random set of weights, the results of training can vary. We report the median score after 10 train test cycles. Our results for real world test is given in 3.4.

The results here are very interesting. The most notable fact is that PU encoding helps the conventional LDR NR-IQA algorithms to generalize better. This is not surprising as PU encoding was designed to convert luminance values to a dynamic range independent space. By itself, BRISQUE, BIQI, SSEQ and DIIVIINE seem to be unable to adapt to the different image sizes and luminance ranges in the data set when these are different from the training set.

This can be explained by the fact that the features used by these algorithms are computed over a joint histogram from the entire image.

The kCNN method performs well and shows good adaptability to a different test case. This can be attributed to the fact that an image block is used to train the kCNN and hence the overall image size becomes less important. The method, however, does not take into account perceptual factors and we see an improvement if PU encoding is used.

Our proposed algorithm outperform all the test cases by a large amount in this test of generalization to real world scenario. The large differences in performance shows the strength of the two-stage method. The proposed model successfully adapts to different image and luminance range of distorted images in test conditions because of the perceptual Resistance values scaling the error in accordance with the luminance and the content. The method is achieves performance close to full reference algorithms, though there is still room for improvement. Note here that the large RMSE value for full reference algorithms are because the scale of the values produced by these algorithms are significantly different compared to original DMOS.

A scatter plot of the scores produced by the proposed method to actual DMOS is shown in Fig 3.8

3.5.6 Error Estimation

The output of E-net after real world test on a few images from Data set #4 and #5 are shown in Figure 3.9. The figure shows a few distorted images, the error present in those images and the estimation of error by E-net. We colormap the images for easier visualization; red represents high values, green intermediate and blue low values. The corresponding Mean Squared Error (MSE) between the actual error value and estimated error value is given in the last column. Note here that the errors are being compared here and not perceptual quality, hence the use of MSE. It is clear that E-net performs as expected and is able to successfully isolate most of the errors in the image.

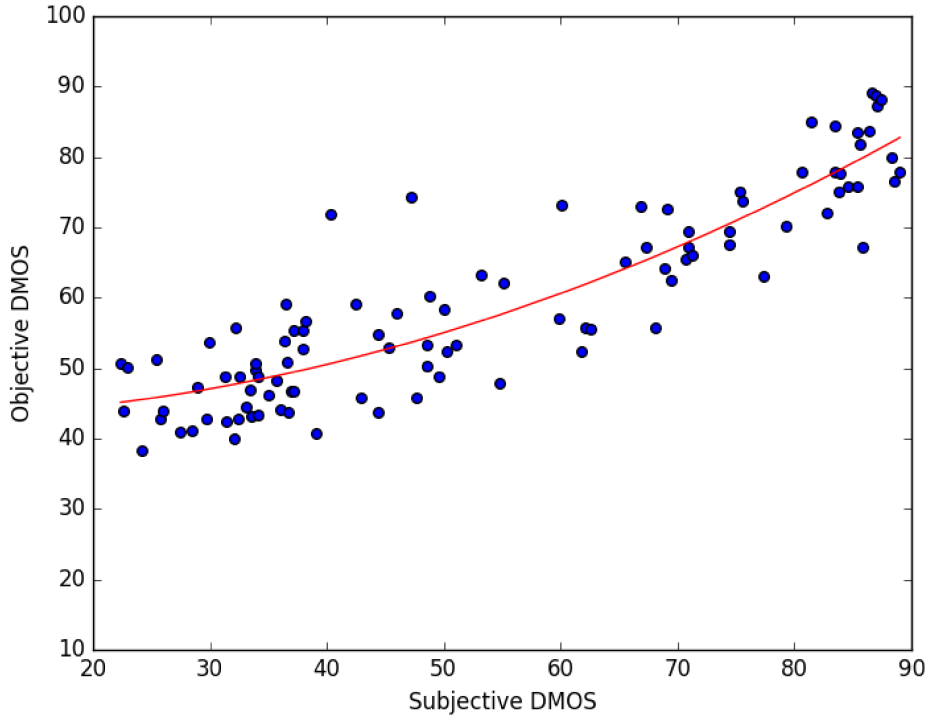


Figure 3.8: Scatter plot between objective scores by proposed method and actual DMOS.

3.5.7 Perceptual Resistance

The perceptual component of the proposed architecture is the Perceptual Resistance produced by P-Net. We expect this value to change depending on the image data. A large value of luminance and high frequency would generally have a larger masking effect leading to low error visibility and hence we expect the Perceptual Resistance to be high. Conversely, low and medium brightness with slow variations would be attributed to lower masking value, higher error visibility, and less Perceptual Resistance.

Demonstration of masking effects

To further demonstrate the characteristics of the output of P-net, we consider the image in Fig 3.10. The image is compressed using JPEG compression with a Q factor of 10. The distorted image luminance value is linearly scaled to have a maximum values of 4000 cd/m^2 , 2000 cd/m^2 , 1000 cd/m^2 , 500 cd/m^2

Feature	Processing	SRCC	KRCC	PLCC	RMSE
BRISQUE	Lin	0.5400	0.3732	0.4772	28.8475
	PU	0.7135	0.5121	0.6503	20.5534
	TMO - Drago	0.6337	0.4483	0.5903	21.7118
	TMO - Reinhard 02	0.6583	0.4670	0.6512	18.4500
	TMO - Reinhard 05	0.3524	0.2482	0.3946	30.6615
	TMO - Mantiuk	0.5887	0.4103	0.5493	22.7529
SSEQ	Lin	0.5287	0.3599	0.4714	25.2588
	PU	0.6492	0.4543	0.6111	19.6977
	TMO - Drago	0.5865	0.3956	0.5634	22.6987
	TMO - Reinhard 02	0.5810	0.4075	0.5644	22.9900
	TMO - Reinhard 05	0.4990	0.3401	0.5036	24.9193
	TMO - Mantiuk	0.4973	0.3398	0.4770	21.2044
BIQI	Lin	0.2845	0.1876	0.2831	31.0686
	PU	0.4386	0.3041	0.4399	21.2084
	TMO - Drago	0.5332	0.3780	0.4436	25.6200
	TMO - Reinhard 02	0.4632	0.3196	0.4358	22.0376
	TMO - Reinhard 05	0.5748	0.4048	0.5630	19.4825
	TMO - Mantiuk	0.4651	0.3204	0.4571	24.2268
DIIVINE	Lin	0.5041	0.3429	0.5209	20.6506
	PU	0.5318	0.3691	0.5442	19.6772
	TMO - Drago	0.4143	0.2852	0.4065	25.9697
	TMO - Reinhard 02	0.3634	0.2434	0.3953	26.1464
	TMO - Reinhard 05	0.5558	0.3849	0.5374	19.3122
	TMO - Mantiuk	0.4138	0.2838	0.4496	21.0499
kCNN	Lin	0.6991	0.5156	0.7008	19.3677
kCNN	PU	0.7694	0.5845	0.7544	18.5854
Proposed	Lin	0.8672	0.6773	0.8780	18.6268
HDR-VDP-2.2	Full Reference	0.9298	0.7691	0.8710	16.2727
HDR-VQM	Full Reference	0.9193	0.7444	0.8940	51.1045
PU-MSSIM	Full Reference	0.8969	0.7125	0.7589	59.7094
PU-SSIM	Full Reference	0.9121	0.7339	0.7121	59.7065

Table 3.4: Generalization capability.

and 250 cd/m². We then give this image as input to the proposed algorithm (as blocks of size 32x32) and examine the output images. The results are shown in Fig 3.11. The luminance increases from left to right; The distorted image is the same. The results from various stages processing in the proposed algorithm is shown. The DMOS scores generated by the proposed algorithm is shown in the last row.

Spatial masking: To analyze the spatial masking trends observed in the outputs of P-net (Row 3), consider any one Perceptual Resistance map of one luminance range. The value of the Perceptual Resistance is low in pixels corresponding to the sky (blue values in Perceptual Resistance map), indicating an increase in sensitivity to error and hence a reduced quality. On the contrary, the regions where there is texture, like the bushes or the struts of the tower,



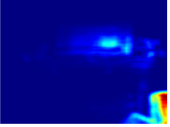
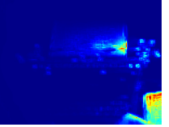


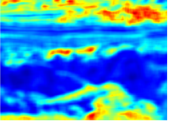
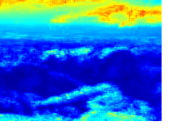


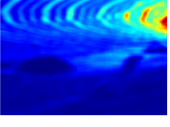
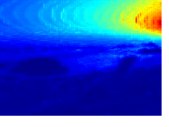
Reference Image	Distorted image	δ	$\hat{\delta}$	RMSE
				4.176
				3.489
				8.018
(a)	(b)	(c)	(d)	(e)

Figure 3.9: Left column distorted images, second column ground truth error, third column the error estimation results from E-net and final column shows the MSE between Estimated and Actual errors on the HDR luminance image.

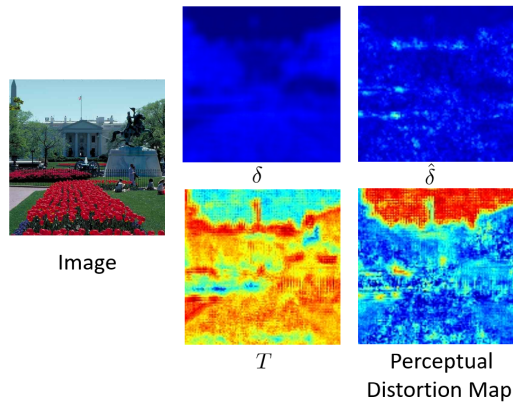


Figure 3.10: Image for studying the nature of Perceptual Resistance.

the Perceptual Resistance is higher indicating less susceptibility to error.

Sensitivity change with luminance: As explained in III.A, [7] did a user study on a distorted image being displayed in two screens with maximum luminance 300 cd/m^2 and 1000 cd/m^2 . The study revealed that users rated the quality of distorted image being displayed on high luminance monitor as worse compared to the the same image displayed on the lower luminance monitor.

We see the similar results in our model. Referring to Fig 3.11, the DMOS predicted is increasing with luminance; implying a reduction in quality and an increase in error visibility in high luminance displays. Thus we indepen-

dently confirm the findings of [7] using a pure data driven approach without psychophysical measurements.

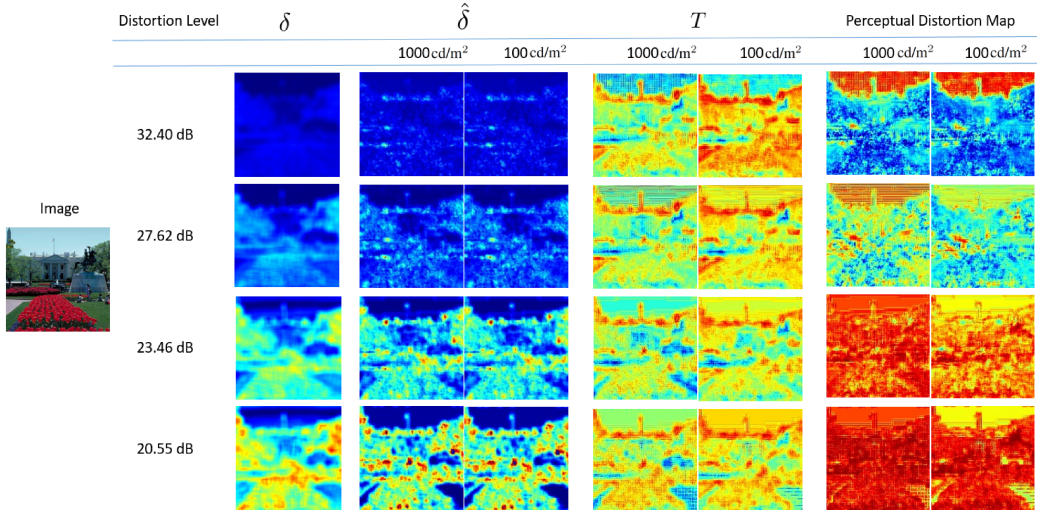


Figure 3.11: Input image (top row), Estimated error by E-net (second row), Perceptual Resistance by P-net (third row) and the local error map from the mixing function (fourth row) for different luminance ranges obtained by linearly scaling image intensities in range [0,4000]. Red implies high value and blue implies low.

The change in DMOS is caused entirely by the perceptual resistance values generated by the P-net. As shown in the figure, the Estimated error is constant regardless of the intensity scaling of the input image. This proves that our system captures the perceptual sensitivity changes associated with changes in luminance.

Comparisons with perceptual models

The only algorithm that provides a perceptual threshold for HDR is HDR-VDP2.2. The contrast threshold values produced by HDR-VDP2.2 are similar to Perceptual Resistance values, even though the final prediction in HDR-VDP2.2 is a probability of error detection and ours is DMOS. This comparison is shown in Fig 3.13. The values shown in the figure are in log scale normalized to span [0,1]. We do this because a one-to-one comparison between the two results would not hold any meaning due to the differences in range and scaling of the values. This result clearly shows the similarities in perceptual results

of the two algorithms.

Distorted Image	δ	HDR-VDP 2.2	PU-SSIM	PU-KCNN	Perceptual Distortion Map	MSE with HDR-VDP	
						Proposed	PU-KCNN
						0.0981	0.132
						0.1450	0.2070
						0.2446	0.3958
						0.1536	0.1834

Figure 3.12: Comparison of distortions in image or error maps estimated by various IQA schemes. Red represents high value and blue represents a low value.

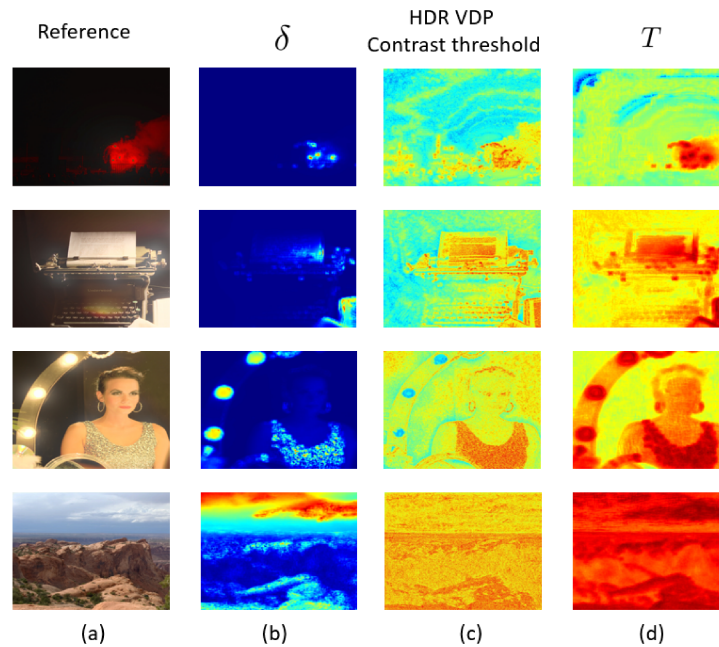


Figure 3.13: Comparison of the output of P-net with HDR-VDP2.2 contrast threshold. Image pixel values are in log scale and normalized.

3.5.8 Error maps

One of the advantages of a CNN based NR-IQA scheme is that it gives the approximate location of the perceived errors. We refer to this as the *error map*.

PU-SSIM and HDR-VDP2.2 uses Minkowski summation, whereas we use a mean value for the final quality. However, a relative comparison is helpful to know which areas of the image show errors. Again the results we report here are on data set #4 and #5 after the test for generalization capability.

A comparison of the error maps produced by HDR-VDP2.2, PU-SSIM, PU-KCNN and the proposed method is shown in fig 3.12. In the figure, for HDR-VDP2.2, the probability of error detection is shown; for PU-SSIM, an inverted PU-SSIM map is shown to indicate the areas with error as high values. The output of proposed algorithm is shown in the fourth column. We normalize all the values so that the maximum value is one in which we get a good relative comparison. The images are color-coded - red represents high value, green intermediate values and blue represents a low value. The error maps produced by proposed scheme and kangCNN with PU processing are compared with HDR-VDP2.2 and corresponding MSE value are reported.

It is clear that the values produced agree with the highest performing full reference metric (HDR-VDP2.2) in terms of the location of and the relative intensity of the visual errors. Further proof of the overall performance is can be seen in the high correlation values of proposed metric scores compared to PU-KCNN in Table 3.4.

3.5.9 Effects of mixing function

The nature of P-net is heavily dependent on the specific mixing function. As explained, the choice of this function can be arbitrary as long as it is monotonically increasing and the network can converge.

To show the effects of using alternate mixing functions to the one we proposed to use, we perform the real world test using the following mixing functions.

1. Proposed model,

$$DMOS(i, j) = 1 - \exp\left(-\left|\frac{k*\delta(i, j)}{T(i, j)}\right|\right).$$

2. Linear mixing, $DMOS = \frac{\delta}{T}$.

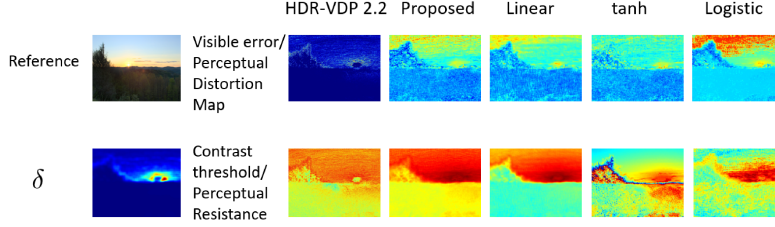


Figure 3.14: Comparisons of results of mixing function (Error map) and P-net (Perceptual Resistance) when different mixing functions are used for training the system. HDR-VDP probability of error detection and contrast thresholds are shown in first column for reference.

3. Hyperbolic tangent, $DMOS = \tanh(\frac{\delta}{T})$.

4. Logistic function,

$$DMOS(i, j) = \frac{1}{1 + \exp(-k(x - x_0))}$$

where $x = \frac{\delta}{T}$.

We found train test cycles for cases 2 and 4 above, where the network failed to converge and the training error kept increasing. This happens when the random weight initialization causes the results of these functions to go to unbounded values and when the values get saturated interfering with the gradient propagation in the CNN. The *nonconverging epochs for cases 2 and 4 were removed* as it does not reflect the performance of the functions. We did not have problems with convergence in case 1 and 3. The results are shown in table 3.5. We see similar performances whenever the network converged, reinforcing the argument that the choice of $G()$ is arbitrary.

To further investigate the results, we select a distorted image and show the results from different mixing function and P-net respectively in Fig 3.14. As explained earlier, output of mixing function represents the DMOS for the image block and that of P-net represents the Perceptual Resistance of the block. A comparison of HDR-VDP2.2 probability of error detection and contrast thresholds are shown for reference.

With respect to the mixing function, the error maps do not change significantly as the mixing function changes. This is because the optimization

	Proposed	Linear	tanh()	Logistic
SRCC	0.8672	0.8616	0.8560	0.8476
KRCC	0.6773	0.6630	0.6719	0.6474
PLCC	0.8780	0.8597	0.8688	0.8535
RMSE	18.6268	18.8270	16.2990	26.7700

Table 3.5: Performance with various mixing functions.

process of the CNN tries to minimize the difference between mixing function results and real DMOS. Once, converged, the results will be similar.

Considering P-net, we see that the results show a similar trend in terms of relative values. Thresholds in the sky are generally higher compared to the texture-rich forest area. However, the scale of change of Perceptual Resistance is different depending on the type of mixing function used. The results of the mixing function defined in equation 3.3 are the closest to that of HDR-VDP2.2.

3.5.10 Failure cases

We found that the algorithm fails in some cases. Examples are shown in fig 3.15.

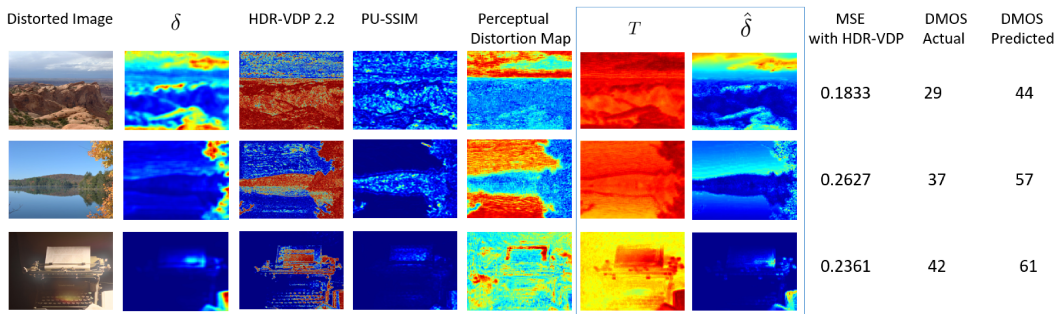


Figure 3.15: Cases where the error maps produced by proposed method fails.

Here we show the error maps for proposed method, HDR-VDP and the corresponding error estimation and Perceptual Resistance images produced using our method. We observe that the error maps produced by our method do not correspond to the ones produced by HDR-VDP2.2 and PU-SSIM. This is especially true in the sky region where both full reference algorithm claim

that there is minimal distortion, however, our method predicts a higher error. MSE values are given to quantify the results.

To investigate the issue further we show the corresponding error estimate and the perceptual resistance values in columns 5 and 6. The error is caused by faulty error estimation by E-net and these faulty value propagating to P-net. The training process creates an internal bias that produces error values that tend to be always high in smooth regions. This is most likely due to the large amount of sky pixels in training data affecting the training.

3.6 Conclusion

We propose an HDR NR-IQA scheme that uses a CNN based architecture to generate values corresponding to perceptual masking and true Error present in an image and combines it in a mixing function. The perceptual effects are derived from optimization on real world data and do not involve psychophysical measurements. The perceptual resistance derived from data show similar characteristics as other perceptual models. Our algorithm predicts the visual distortions in the image due to low-level distortion such as compression artifacts. It was found that the algorithm scores correlate well with human scores. It outperforms state-of-the-art NR-IQA methods and is competitive when compared to HDR FR-IQA methods.

Chapter 4

Learning error visibility from Image quality

4.1 Introduction

A visual error detection threshold measures, the magnitude a certain target stimulus must have, in order to become distinguishable from a background, masking signal. They are useful to determine the error visibility of various kinds of distortions (compression artifacts, additive noise, etc.) produced by several image processing algorithms. Predicting the visibility of visual distortion is of paramount importance in a number of image processing applications, such as image compression, watermarking and quality assessment.

Conventional approaches to model distortion visibility strongly rely upon psychophysical experiments that are, in their nature, based on a simplification of real-world conditions. For example, models that describe visibility of sine-wave gratings might be enough to predict visibility in DCT-based image compression; however, they are probably failing in modeling different or multiple concurrent artifacts. Furthermore, local visibility is influenced by surrounding regions, and is ultimately linked to image semantics. It is evident that modeling all these complex factors *only* through psychophysical experiments is unfeasible.

4.2 Motivation

Instead of learning distortion visibility directly from psychophysical data, we propose to learn it *indirectly*, leveraging the large availability of alternative, yet related, data: subjectively annotated image quality assessment (IQA) datasets. Image quality scores provide higher level information about the visual appearance of a picture, compared to psychophysical measurements. At the same time, they bring information about the visibility of distortion. Indeed, a common assumption in image quality assessment is that the perceived quality is directly related to the visibility of the error signal [109], [129]. In other words, the per pixel error is weighted locally by the ensemble of perceptual phenomena, such as contrast sensitivity and several forms of masking, which discount its visibility to the human visual system.

Visual quality scores thus implicitly embed latent information on error visibility. In Chapter 3, we had proposed a deep convolutional neural network (CNN) architecture to disentangle the per pixel distortion and what we called the “perceptual resistance”, in the context of no-reference quality estimation of high dynamic range compressed pictures. Our results demonstrated that it is possible to effectively estimate these two terms over a broad range of qualities, starting from supra-threshold quality scores. This opens a question of whether a similar approach can also be used to predict near-threshold visibility. To this end, we train our proposed system in a full-reference fashion, i.e., we assume that the error signal is known, as it is the case in most IQA datasets. However, the inference step does not require the knowledge of the error, and can produce an estimate of local masking for any input image. Interestingly (and perhaps surprisingly), we find that perceptual scaling learned from image quality scores can predict the detection thresholds in [3] with similar accuracy as the CNN-based regressor in [4], although our model is learned on other datasets with different contents and several kinds of visual impairments. This makes the proposed approach potentially more general than previous work.

4.3 Proposed model

In this section we present our proposed model to estimate local distortion visibility thresholds. Specifically, we first discuss the assumptions of our model and provide a mathematical framework to split (supra-threshold) quality scores into per pixel error and a perceptual scaling term, which we show in Section 4.4 to be a good predictor of visibility thresholds in local masking. Afterwards, we describe how to implement this model using a deep convolutional neural network architecture.

4.3.1 Mathematical framework and assumptions

Let I_R and I_D be an original reference image and its distorted version, respectively, and $Q \in [0, 1]$ the quality score for I_D . Without loss of generality, we assume that Q is given as a differential mean opinion score (DMOS). We assume that we have access to the local quality $q(i, j)$ of an image patch $I_D(i, j)$ of size $N \times N$ pixels, centered at location (i, j) . The pooling process linking the quality of individual patches to the overall quality Q may depend on many factors, e.g., saliency. Here, we assume that the local quality of an image block is the same as the global image quality score, similarly to the setting in [44]. While this is a strong assumption, which is often not met in practice, it has been proved to be accurate enough to predict image quality [44].

In order to model local quality, we further assume that per pixel distortion in a patch is discounted by a perceptual weight, $T(i, j)$, that accounts for typical masking and visibility effects. Specifically, we measure pixel distortion through the average absolute error $E(i, j)$ of a patch, defined as:

$$E(i, j) = \frac{1}{N^2} \sum_{k=1}^{N^2} |I_D(i, j)_k - I_R(i, j)_k|, \quad (4.1)$$

where k is the pixel index in the patch. We then approximate local quality as a function of the error and the perceptual weight, that is:

$$q(i, j) \approx 1 - \exp \left(- \left| \frac{\alpha \cdot E(i, j)}{T(i, j)} \right|^\beta \right), \quad (4.2)$$

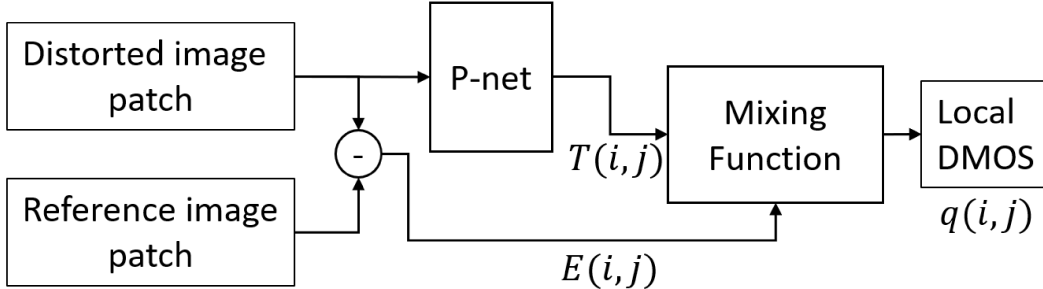


Figure 4.1: Neural network architecture for extracting contrast detection thresholds from IQA databases.

where α and β are two model parameters.

Notice that this formulation has been previously used to model the probability of detecting localized noise distortion, e.g., in [129], but we introduce an additional scaling factor α . Eq. (4.2) is inspired by the common practice in vision science of expressing the magnitude of the error in multiples of the just-noticeable difference [113], although the relationship in this case is nonlinear. We therefore refer to $T(i, j)$ as the *local visibility threshold*, even if this is technically correct only for near-threshold distortion. We show in Section 4.4 that, for this latter case, T does indeed model local visibility. Moreover, the thresholds we estimate are not in the same scale as those obtained through psychophysical experiments. We compensate for this by means of the parameter α , which provides an additional degree of freedom in the optimization to push the predicted local quality as close as possible to ground truth. The value of the parameter β is generally found by matching the results of psychophysical experiments. In practice, we found that the choice of β does not significantly affect the performance of our model, and α alone provides already enough flexibility to minimize the loss function. Hence, we simply set $\beta = 1$ in Eq. (4.2).

4.3.2 Implementation

A scheme of the proposed model is shown in Figure 4.1. The input of our systems are overlapping patches of 32×32 pixels, similar to [44], [53]. The distortion, $E(i, j)$, is computed directly as in Eq. (4.1) during the training,

where both I_D and I_R are available. The local visibility thresholds $T(i, j)$ are computed in a module that we name “P-net”, while an estimate of local quality, $\hat{q}(i, j)$ is obtained by implementing Eq. (4.2) in the “Mixing function” block. Notice that this structure is required for training the P-net, as T is considered a latent variable which depends implicitly on the observations of the input content and perceived quality. Instead, for inference the P-net is employed as a standalone block. Furthermore, we are generally interested in applying the learned P-net on the *original* pictures, rather than on the distorted ones. However, we found that training the P-net with noisy versions of the image was more effective, as this increases the variability of input data, leading to improved generalization capabilities.

Our model is trained to predict local quality $\hat{q}(i, j)$ using the ground-truth quality $q(i, j)$ as targets, by minimizing the following cost function:

$$J(i, j) = |q(i, j) - \hat{q}(i, j)|. \quad (4.3)$$

Notice that $\hat{q}(i, j)$ depends implicitly on the latent variables $T(i, j)$ through Eq. (4.2). Thus, when optimizing J , the visibility thresholds are adjusted in such a way to weigh the error coherently with the observed ground-truth quality.

For the architecture of P-net, we make use of a handcrafted layer that we named as augmented input layer [53]. In this layer, in addition to the luminance values of the $N \times N$ block, we compute the mean, variance and Mean Subtracted Contrast Normalized (MSCN) image [77]. The latter is defined as:

$$MSCN(x, y) = \frac{I(x, y) - \mu_M[I(x, y)]}{\sigma_M[I(x, y)] + \epsilon}, \quad (4.4)$$

where (x, y) denotes the location of a pixel in the patch, $\mu_M[I(x, y)]$ is the mean and $\sigma_M[I(x, y)]$ the variance of the patch, computed by replacing every pixel (x, y) with the mean and variance, respectively, over a local Gaussian window of size $M \leq N$ around (x, y) . The regularization term ϵ is set to 0.01. This is followed by convolutional layers with $32 \times 5 \times 5$ filters and a fully connected layer with 100 nodes. We use *relu* activation in all neurons. Dropout layers are used to prevent overfitting.

4.4 Results and Analysis

In order to assess how well the estimated distortion visibility predicts ground-truth data from psychophysical experiments, we test the proposed model on the dataset of local masking thresholds in [3]. This dataset collects measured threshold values for 1080 image patches of size 85×85 pixels, extracted from the CSIQ dataset [59]. The detection thresholds are reported in terms of root-mean-squared (RMS) contrast and expressed in decibels (dB).

To test the proposed model, we train it on three different datasets: the CSIQ dataset [59], containing 855 images, 6 types of distortions; the TID 2013 dataset [90], with 3000 images and 25 types of distortion; and the LIVE dataset [37], featuring 779 images and 5 types of distortion. In general, the thresholds $T(i, j)$ found with our model do not lie on the same scale as those in [3]. In addition, the P-net can be trained on different IQA datasets, and the interpretation of DMOS in each dataset depends on the experiment carried out to collect the data [123]. Following a typical protocol in the evaluation of quality metrics, we compensate for this mismatch by linearizing the predictions with respect to psychophysical ground truth through a monotonic third-order polynomial fitting before evaluating their statistical accuracy.

4.4.1 Performance

A comparison of the predicted and ground-truth local visibility thresholds is illustrated in the scatter plots of Figure 4.2, where our P-net has been trained on the LIVE and TID 2013 datasets, respectively. Each point in the scatter plot represents a 85×85 patch of [3]. Since our predictor can produce per pixel estimates of distortion visibility (using overlapping patches), we decimate the maps produced by the P-net to match the resolution of the ground truth, using a simple averaging filter (see Figures 4.4 and 4.5).

We can observe from Figure 4.2 that the predicted thresholds capture relatively well the overall trends of the measures obtained from psychophysical experiments, even if they have been obtained by training on very different data (IQA scores) and using different source contents. This indicates that learning

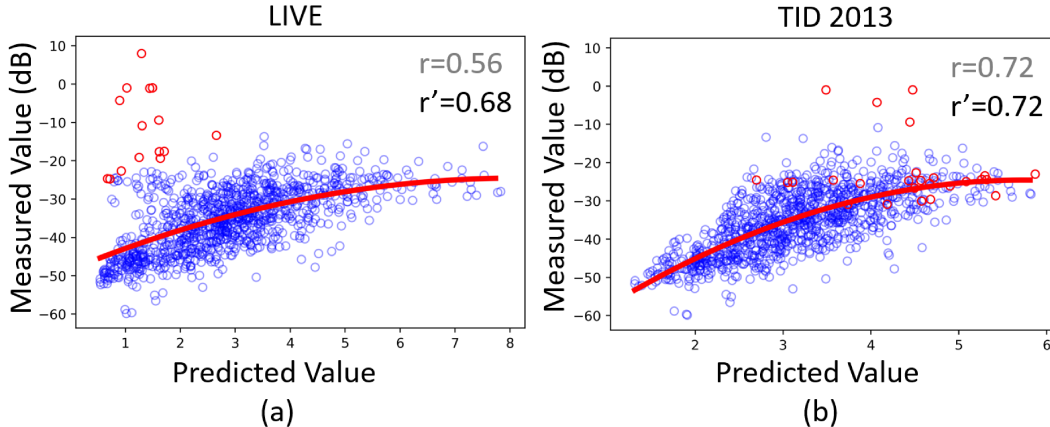


Figure 4.2: Scatter plots of contrast detection thresholds derived using our method vs experimentally measured values. The system is trained on (a) LIVE dataset (b) TID 2013 dataset. The polynomial fitting line is shown in red. Red points correspond to patches whose luminance is outside the range of the training datasets, see Figure 4.3. r is the PLCC (after fitting) on the whole test set; r' is the PLCC excluding the red points.

visibility thresholds from generic IQA datasets is feasible and can generalize sufficiently well. Nevertheless, we notice in Figure 4.2 that in some cases the predicted thresholds deviate significantly from the measured ones. Especially for the LIVE dataset, this degrades the performance, measured by the Pearson linear correlation coefficient (PLCC) $r = 0.56$. To further investigate this, we analyze the distribution of the average luminance intensity of patches in the TID and LIVE datasets compared to CSIQ in Figure 4.3. Notice that there is only a negligible amount of patches for the LIVE dataset in the intensity range $[0, 10]$ and $[250, 255]$. This lack of data can affect the performance of prediction in this specific intensity range. We verify this by highlighting in red those patches of the test set having luminance outside the interval $[10, 250]$ in Figure 4.2. We observe that these points correspond indeed to the outliers in the scatter plot. By removing these few very dark or very bright patches (28 patches out of 1080), we observe that the PLCC increases to $r' = 0.68$.

We compare in Table 4.1 the performance of our method (trained with three different IQA datasets) with other predictors of local visibility thresholds proposed in the literature. The majority of these methods use handcrafted

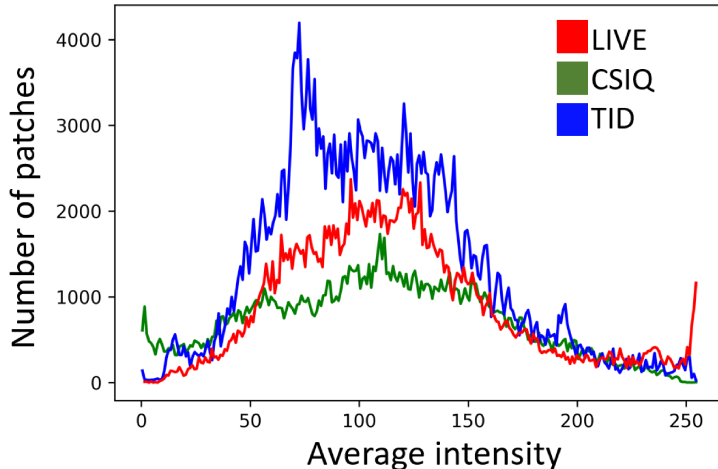


Figure 4.3: Distribution of average intensity of 32×32 patches in LIVE, CSIQ and TID datasets. The distributions of the three datasets overlap only in the intensity range $[10, 250]$.

models directly derived from psychophysical experiment, while [4] employs a convolutional neural network. The performance criterion is root-mean-squared error (RMSE) between the ground-truth and predicted thresholds. We observe that our approach provides comparable or better results to state-of-the-art methods. It should be noticed that the two methods in [4] are trained on the same dataset [3] used for test, and thus their performance represents a sort of accuracy upper bound. Conversely, the proposed method achieve competitive results even when it is trained on the TID 2013 or LIVE datasets.

A qualitative illustration of the predicted thresholds is reported in Figure 4.4, where we also show per picture PLCC. We observe that the perceptual thresholds produced by our approach are intuitive, e.g., thresholds are lower for relatively smooth regions (sky) and higher for more complex regions (grass, leaves). There are of course also cases in which the prediction fails (see examples in Figure 4.5). This mainly happens in dark patches, as discussed above, where the training datasets do not offer sufficient samples to learn robustly the visibility thresholds.

Method	Training Data	RMSE
Watson <i>et al.</i>-KMF [112]	Psychophysical visibility experiments	5.713
Watson <i>et al.</i> -JYS [112]		6.521
Teo & Heeger [102]		6.861
Chandler <i>et al.</i> [13]		6.879
Optimized GC [4]		5.192
Alam <i>et al.</i> CNN [4]		5.475
Proposed		CSIQ quality scores [59]
	LIVE quality scores [37]	5.991
	TID 2013 quality scores [90]	5.626

Table 4.1: Performance comparison between different algorithms. Highlighted are the best state-of-the-art methods (handcrafted and CNN-based), as well as our results.

4.5 Conclusion

We present a method to derive local visibility thresholds from image quality scores using a neural-network-based approach. Our experiments demonstrate that the latent information about distortion visibility carried by supra-threshold quality scores can be recovered and used to predict near-threshold local masking. One advantage of our approach, compared to models based on psychophysical data, is that it can leverage the larger availability of subjectively annotated image quality datasets. We plan to formalize further this approach in the future and apply it to objective image quality assessment.

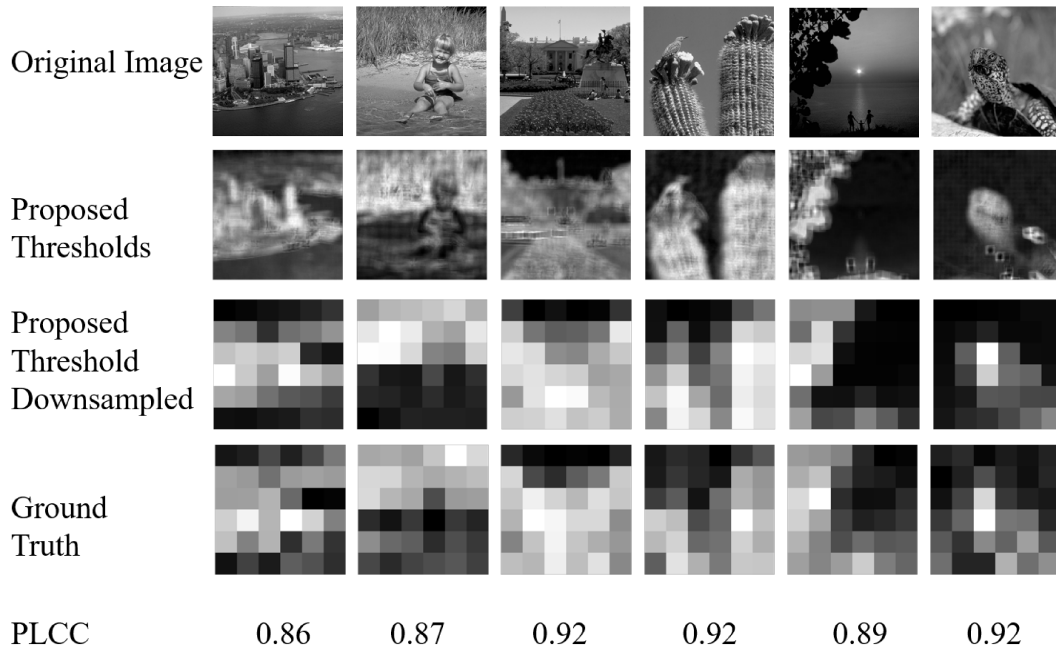


Figure 4.4: Examples of local visibility thresholds produced by our method trained on TID 2013. The estimated thresholds are decimated to match the resolution of the ground truth. Both sets of values are scaled to fit in the range $[0, 1]$ for visualization, with 0 (black) being the lowest threshold.

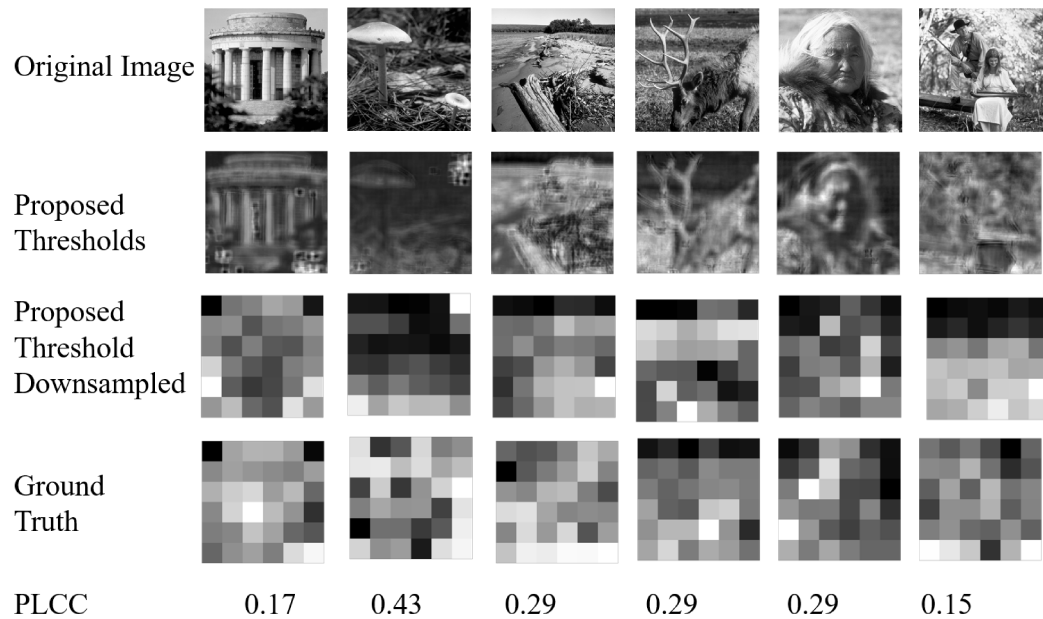


Figure 4.5: Some failure cases of the proposed method trained on TID 2013. Often these cases correspond to patches with low luminance, which are under-represented in the training data.

Chapter 5

Content dependency

5.1 Introduction

In the Human Visual System (HVS), the quality evaluation process is carried out at different layers, equivalent to starting from pixel-level to neighbourhood, and to scene, in image processing and pattern recognition. An explanation of how high-level scene information affects low-level processing can be found in [34]. The authors showed that low-level feature maps can be preset at higher-level and influence in a top-down manner. In other words, if the viewer is familiar with the scene, expectation creates certain degree of influence when interpreting low-level features. Thus, depending on the viewer's expectation of the scene, his/her visual quality score changes. A simple example is while enjoying a natural scene image, e.g., rural landscape under the sun, we expect a blue sky, fresh green leaves, bright colored flowers, etc. The deviation from a blue sky, green leaves and colorful flowers affects the perceived image quality. Conversely, if we judge a portrait image, we look for vividness of the eyes, correct skin tones and so on. Any deviation affects the image quality.

Based on the above real-life experience, a practical approach to assign the degree of importance of an image region relating to visual quality is a scene dependent Visual Error Importance (VEI) map, where the values in a region reflect the influence to visual quality. We hypothesize that if scene characteristics is given as prior knowledge, the values of this map can be computed by a second order function with low-level image features as parameters. The parameters can be derived by using a global optimization process and an Im-

age Quality Metric (IQM). Such a framework can be used to augment existing IQMs. In addition to augmenting the performance of IQMs, VEI can also be used to propagate high-level information (top-down) in various perceptual quality dependent image processing techniques, like image compression and transmission.

The contributions of this work include: 1) Proposing a content-specific IQM performance augmentation strategy, and 2) validating the strategy with a benchmark dataset.

5.1.1 Psycho-visual experimental evidence

A major experimental evidence supporting our proposed strategy was provided by [101]. The study analyzed influence of scene categories on image visual annoyance. Specifically they selected: indoor, outdoor natural, and outdoor man-made. The study showed that humans recognize these scene categories differently at pre-attentional stages. The impact of a scene category on the impairment level was computed by using a Generalized linear mixed model.

They found *statistically significant results showing that a scene category had an influence on the degree of visual annoyance* perceived by the viewer. Specifically, they found that people are more critical to indoor images as compared with outdoor images, especially when comparing with outdoor natural scenes. Another observation in the experiment was the difference in quality assessment, which seems dependent on scene category.

In a related study [101], the authors observed that there was a difference of importance between non-animated and animated objects. Their experimental results indicate that object categories in the image also have effects on perceived quality. Viewers tend to be more critical towards images with animated objects rather than non-animated (man-made or not alive) objects.

Further insight on how image quality is influenced by scene features and high-level content was provided by [5]. Here, the authors found category-specific dependence of colorfulness on aesthetics. A main conclusion of the study was that though there was no direct correlation between colorfulness and image quality, if the image was macro photograph, colorfulness was found

to correlate with visual quality.

Motivated by previous studies including psycho-visual experimental evidence, we conducted a test described in the next section to verify our hypothesis of "Visual importance in an image can be influenced by the scene category."

5.1.2 Experiment on CSIQ

A quick verification to find out whether image quality is content-dependent can be obtained by analyzing the impact of noise on images with natural and man-made scenes. Sample images are available in the CSIQ dataset [59]. We choose images that was clearly outdoor man-made (not alive) and clearly natural (alive) (Fig. 5.1), and analyzed the Difference in Mean Opinion Scores (DMOS) for all the distorted images in the dataset. The amount of noise remained the same for both natural and man-made content. We found that the overall mean of DMOS for outdoor man-made images was 0.35 and for outdoor natural images was 0.41. The difference was statistically significant ($p < 0.05$). It means that the HVS perceives higher quality in images with man-made scene compared with natural scene, given the same noise distortion level. In other words, image content influences image quality assessment. Although the difference in score, i.e., 0.06, is small, which is likely due to the Double Stimulus Impairment Scale method used for generating the DMOS, the statistically significant difference verifies the observation of [101], that people are more critical to images with natural animated objects, as reflected from a larger DMOS (poorer perceived quality), and how the HVS assesses image quality is also dependent on the scene category or the objects composing the scene.

5.2 Computational Design and Implementation

From the discussion above, we believe that human perception of image distortions changes depending on the high-level image content. Certain features like color become more visually important in certain specific categorizes of image

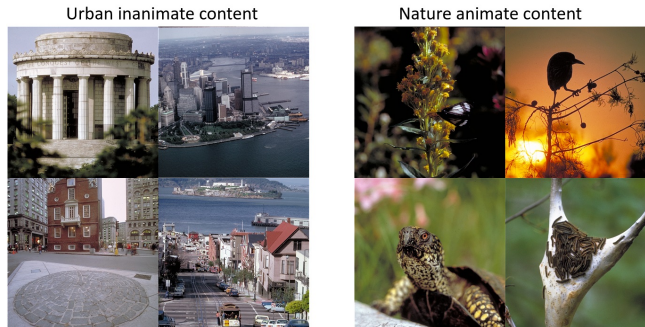


Figure 5.1: Examples of outdoor natural and outdoor man-made images chosen from the CSIQ dataset.

content. In other words, if an Image Quality Metric (IQM) integrates this content-specific factor in its computational model, the image quality assessment performance can be improved.

To explore this strategy, we define a high-level *Visual Error Importance* (VEI) map from a reference image (original image without distortion) and use it to mask the error map of an IQM. The VEI map displays the visual significance of image errors dictated by the image content. Computation of the VEI map is done via an optimization framework using a nonlinear combination of low-level image features of the reference with specific content. Note that VEI computation does not involve local processing of noise and is purely based on image content of the reference.

There are two important parameters in our formulation: image content C and image features F .

5.2.1 Image Content C

For generating vector C , a content detection system needs to be implemented. In this work, we are more concerned about the overall influence of the image and not that of any particular object. Hence, we limit the definition of content to just scene represented in the image. We consider the following broad categories of scene content: 1) indoor: characterized by rectangular spaces like rooms lit by artificial light or low lighting with both man-made and natural objects, 2) outdoor urban man-made: characterized by man-made buildings

Ind: 0.4936, Od-Nat: 0.0107, Od-Man: 0.4956



Figure 5.2: Scene Probability: The description on top shows the likelihood of each class in the image: Id refers to indoor, Od-Nat is outdoor natural, Od-Man is outdoor man-made.

and structures in sunlit or lit by artificial light sources, and 3) outdoor natural: characterized by fresh colors, natural objects like trees or animals, lit by sunlight.

We used the scene-15 dataset provided by [64] and reorganized the images for our scene classes. Then, we used a combination of SIFT [70] and GIST [83] features with a multi-class SVM using RBF kernel to train the scene classifier.

Practically, there can be multiple classes in a single image. Hence, we use the class probability of the SVM. Corresponding to each image, we have a vector C consisting of c_s , where c_s denotes the probability of the image containing a class s . An example image with corresponding scene probabilities is shown in Fig. 5.2, where the scene probabilities generated by our classifier indicates that the scene has both indoor and outdoor man-made characteristics.

5.2.2 Image Features F

We use the Hue, Saturation, Entropy, image Detail level and image Saliency to generate feature maps. To compute the detail level of an image I , we use the equation, $f5(I) = abs(I - G \circledast I)$, where $f5(I)$ is the feature map computed, \circledast is the convolution operation and G is a 3x3 Gaussian kernel with standard deviation 1 (though larger blur levels can be applied, we found that this set of parameters gave adequate results). For entropy, we threshold the maximum entropy at 4.0 and then normalize it. We normalize all of the values to the

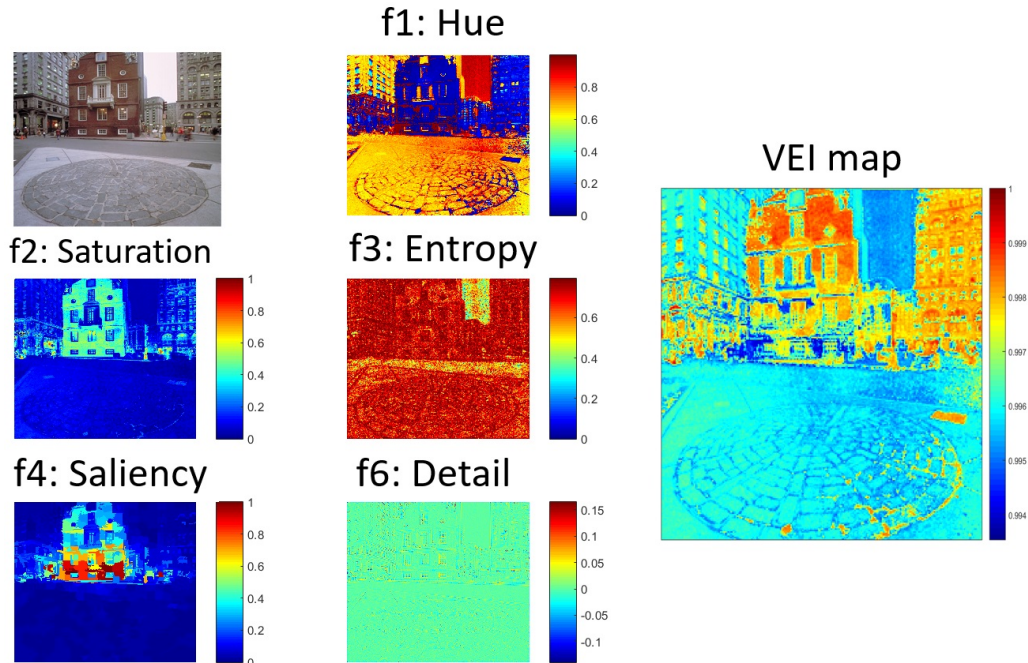


Figure 5.3: Visual presentation of the feature maps extracted from an image. The colorbar with the relative values from cool to hot are shown next to the image. The original image is shown on top left. The Visual Error Importance (VEI) map is shown on the right.

range $[-1, 1]$. These features are used to distinguish different scene content. Natural outdoor images are characterized by bright saturated colors. Outdoor urban environments have rich detail in man-made structures and clear detail in sky. Man-made indoor images usually have less saturated colors, but with salient objects more prominent. We use the saliency algorithms in [49], which adopts a bottom-up approach. The purpose of the saliency map here is to highlight the areas of the image that have an object (or object like structure). An example set of feature maps are shown in Fig 5.3.

5.2.3 Problem formulation

For any distorted image, I_D , let the image content be denoted by a vector $C = [c_1, c_2, \dots, c_M]$, where c_i denotes the probability of the image representing a specific content i . Let image I_D be associated with a set of features represented by $F = [f_1, f_2, \dots, f_N]$, where f_i represents the i^{th} feature map in F . Each feature map has the same resolution as the error map $\delta_{IQM}(I_D)$, of the IQM computed

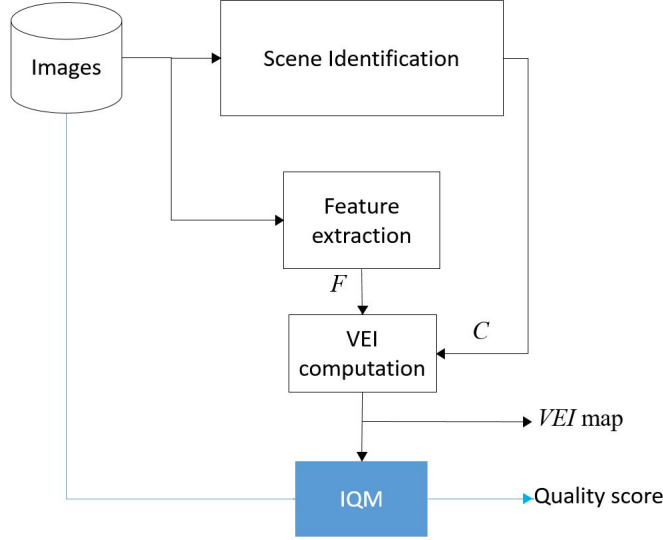


Figure 5.4: Schematic representation of the proposed strategy.

on image I_D .

We compute VEI of I_D as a weighted second order function of all the feature maps in F .

$$VEI(I_D) = \sum_i^N \sum_{j=0}^N w_{i*N+j} * f_i * f_j + \gamma \quad (5.1)$$

where $VEI(I_D)$ is the scaling applied to $\delta_{IQM}(I_D)$, w_{i*N+j} represents the weight of the term $f_i * f_j$, f_i and f_j represent the i and j th feature maps, i.e., $f_0 = 1$. γ is a very small constant used to stabilize the cost function when $w_i = 0 \forall i$. This formulation represents a generic function that can apply to any first order and second order combination of the feature maps. Note that we assume the third and higher order interaction between the features are insignificant because the feature map values are normalized to a maximum value of 1.

Every w is then computed using an optimization process. We simplify Eqn. 5.1 by combining $f_i * f_j$ and $f_j * f_i$ to a single term. The number of terms in w is equal to the number of unique combination of j and k ; let this count be L .

The weight w can vary based on the scene content. Hence for M scene categories, we need $L \times M$ weights. Let this matrix be X . We compute the

weight matrix $W_{I_D} = [w_1, w_2 \dots w_L]$ for an image using Eqn. 5.2

$$W_{I_D} = C * X \quad (5.2)$$

This formulation allows the IQM adjust the weights based on the scene probability, which represents the likelihood that the image contains certain content. If the probability values of a scene are similar to two image categories, e.g., indoor and outdoor natural, the weights would scale and mix accordingly.

We assume that the IQM has the capability to compute the visual quality by pooling the values of its error map. For an image I_D , let the error map be I_{Err, I_D} . Here, VEI gives the top-down influence on the lower-level error map. We apply a linear mixing to simulate the influence of content on the error map, following the prediction technique by [34]. The augmented error map is computed using Eqn. 5.3.

$$\Delta_{IQM, X}(I_D) = \delta_{IQM}(I_D) * VEI_X(I_D) \quad (5.3)$$

where $\delta_{IQM}(I_D)$ is the original error map generated from the IQM, $\Delta_{IQM}(I_D)$ is the augmented error map and $VEI_X(I_D)$ is the VEI computed using the set of weights X .

The final score computation for an image I_D is computed using Eqn. 5.4.

$$Q_{I_D} = pool_{IQM}(\Delta_{IQM, X}(I_D)) \quad (5.4)$$

where $pool_{IQM}()$ represents the original error pooling strategy used by the IQM and $VEI_{I_D, X}$ refers to the VEI computed for image I_D with the weight matrix X .

The optimal value of X can be obtained by minimizing P in Eqn. 5.5.

$$P = -d(MOS, IQM_{VEI_X}) \quad (5.5)$$

Here, $d()$ represents the distance metric used to compare the performance of the IQM, MOS is the ground truth Mean Opinion Score of the set of images considered, and IQM_{VEI_X} is the IQM scores obtained after the enhancement with VEI using Eqn. 5.4 above.

Algorithm	Default performance				Augmented with VEI				Augmented with Saliency map			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
SSIM	0.8763	0.6705	0.7949	0.6419	0.8904	0.7071	0.7994	8.4225	0.8286	0.6231	0.7913	6.1554
FSIM	0.9242	0.7567	0.8047	1.2898	0.9319	0.7698	0.8505	0.6403	0.7523	0.5521	0.7689	1.4757
MAD	0.9466	0.7974	0.9499	6.3578	0.9518	0.8066	0.9446	6.8143	0.9445	0.7952	0.9474	7.2441
GMSD	0.9574	0.8125	0.9336	0.3317	0.9616	0.8232	0.9402	0.4137	0.5485	0.3827	0.6042	0.2627
PSNR	0.5972	0.4208	0.5066	13.219	0.7129	0.5412	0.6510	22.78	0.6066	0.4270	0.5190	9.5203

Table 5.1: Performance of IQMs before and after augmentation with VEI.

In Eqn. 5.5, we constrain the values of X such that $-1.0 < x_i < 1.0 \forall i$. This was done for ease of optimization. The distance metric used was the Spearman Rank Order Correlation (SRCC). We used the technique described in [104] for obtaining the global minimal point. The local optimization was performed by using simplex search method of Lagarias et al. [56].

The overall system implementation with complete optimization is shown in Fig. 5.4.

5.3 Results and Analysis

Our experiments were performed using the CSIQ datasets testing with a number of IQMs: MAD [59], FSIM [126], SSIM [109] and GMSD [118]. Since our strategy is to improve IQM performance using content-specific features, we chose CSIQ because of its rich variation of content comparing to other datasets, e.g., LIVE.

We divide the dataset into an 80:20 ratio based on the image content. Then, we find the coefficient values of VEI using distorted content of the 80% images, and check the performance of the same on the 20%. Note here that we do not alter the selection in any way and there can be some examples of unbalanced training set. The results are obtained after 100 iterations of training and testing. This is to prove that the coefficients we derive using optimization can generalize on a different content. We compare the VEI outcome with the performance of the original IQMs without VEI enhancement (default performance).

Comparison of performance was based on Spearman Rank Order Correlation coefficient (SRCC), Kendall Rank-Order Correlation Coefficient (KRCC)

and Pearson Linear Correlation Coefficient (PLCC). A better IQM performance is characterized by a higher value in SRCC, KRCC, and PLCC. For completeness, we also include Root Mean Square Error (RMSE), which is computed between human score and metric generated score. The improvement in correlation with human judgment is supported by the SRCC, KRCC and PLCC scores.

Comparing the default performance and the enhancement with VEI in Table 5.1, we can see improvement in all the tested IQMs. Note that the improvement in PSNR is large. The degree of improvement varies depending on the IQM used. We can see that the improvement obtained is relatively small. We hypothesize that this is because of the way the IQA experiment was carried out. The MOS scores provided by the CSIQ dataset was generated using the Double Stimulus Impairment Scale (DSIS) method, where the viewer uses an undistorted image as reference and does not need to use any knowledge from experience to judge the displayed images. Hence the top-down influence is less. In a no-reference image quality assessment context, the improved IQM performance can be higher. As shown in our analysis on CSIQ images in section 5.1.2, the IQM enhancement contributed from our content-specific strategy is statistically significant.

Another valuable information provided by our method is the VEI map (Fig. 5.3 Right). VEI maps show the relative importance of visual errors in image regions. We can use this region importance metric to make image compression and transmission algorithms more efficient based on the perceived image content.

It is important to point out that *VEI maps are different from saliency maps* obtained during free viewing as discussed in [33]. The difference can be seen by comparing the outcome obtained by using saliency maps to augment the same IQMs (Table 5.1). For comparison, we applied the saliency map to mask the error map (generated by the IQM) with the equation $\delta'_{IQM}(I_D) = (1 + sal(I_D)) * \delta_{IQM}(I_D)$, where $sal(I_D)$ is the saliency map of image I_D and $\delta'_{IQM}(I_D)$ is the masked error map. We chose this computation as the mixing model for saliency map and error map of the IQM because it can be compared

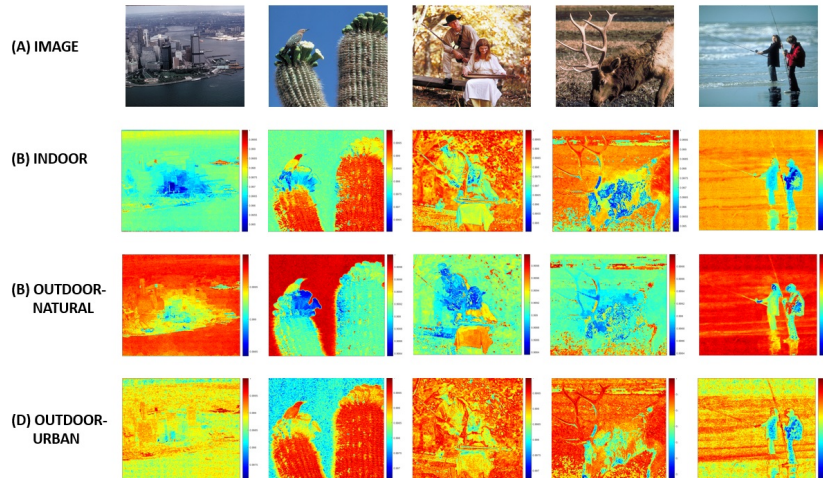


Figure 5.5: We compare the VEI map for the correct class, with the VEI maps generated by forcing the algorithm to consider the image as a different class. Class is shown on the left. The values are normalized to $[0,1]$ to show relative importance. The colorbar on the right of the image shows increasing visual error importance (cool to hot colors).

to our own Eqn. 5.3). We found that if IQM values are reduced to zero by a low saliency, the IQM error can be reduced to a low value. To prevent removal of IQM error, we add a constant 1 to shift the values.

Observed that augmenting with VEI maps, IQMs have better performance than augmenting with saliency maps.

5.3.1 Visual Error Importance (VEI) maps

To gain further insight into how content-specific information can affect our algorithm and output, we force the algorithm to simulate a wrong image content class. Consider the images in the top row of Fig. 5.5 and the corresponding VEI maps in the second to fourth rows when SSIM is used as the IQM. The values in the VEI maps are scaled to the range $[0,1]$ for display purpose. The actual scaling depends on the IQM used.

From the maps, we see that in outdoor natural scenes, VEI emphasizes errors in smooth, well lit areas and lower to medium complexity regions, with a lower priority to salient objects. Such areas are more important if natural scenes are being displayed. On the other hand, for outdoor scenes with man-

made structures, we see that VEI assigns more importance, in term of image quality, to regions with higher details, and less importance to smooth regions. For indoor scenes, we observe a larger importance given to smoother regions.

5.3.2 Limitation and Future direction

Our experimental results and analysis show the feasibility of improving IQM performance by integrating content-specific knowledge. Nevertheless, the effectiveness of the proposed strategy depends on the classifier used to categorize the class content C , the number of classes used and the quality of the features F . All these factors can be handcrafted or incorporated into a machine learning framework to improve performance further. Perhaps a better dataset can be setup as benchmark for researchers to study content-specific influence on IQM.

While the quality assessment scores in CSIQ are based on the Double Stimulus Impairment Scale (DSIS) method, we anticipate to obtain higher IQM improvement if the Single Stimulus Method is deployed. This can be another future direction.

5.4 Conclusion

We propose a new strategy to improve the performance of an IQM by using image content dependent information. Our method produces Visual Error Importance (VEI) maps that detect image regions, which are important in terms of image quality assessment. We tested our proposed method on the CSIQ dataset, which has a rich set of content classes. Experimental results demonstrate that better performances were obtained from all tested IQMs: SSIM, FSIM, MAD and GMSD.

Chapter 6

Color and low-level-feature based quality assessment

6.1 Introduction

My experiments with simple features and HVS based scaling of features was motivated by the fact that most previously proposed theories can be accommodated into a single low level feature based model. In a work I did collaborating with Dr Frederik Dufaux, Dr Irene Cheng and Dr Anup Basu, we exploit an architecture that encapsulates the concept of object detection [42], and explore a feature optimization strategy to deliver a more efficient IQA framework. Current systems often capture edge like structures as features using natural images in the training set. The learnt filters are likely very similar to the Gabor filters that can be used to model the HVS [19]. In this type of model, high level features are generated from the input image. The generated feature representations are used for the object detection process. There are multiple processing layers, with each layer consisting of a filter-bank stage, a non-linearity stage and a feature pooling stage.

Instead of multiple layers, we found out that one level of sub-band decomposition (one layer) with an appropriate feature map scaling based on normalizing the coefficients can accurately predict the effect of image distortion on perceptual quality.

Our low level feature generation approach based on a single-layer object detection architecture is especially effective for IQA because of the following

advantages: 1) the generated features capture structural information [108] of the image, 2) our approach, with a new frequency scaling technique, can explicitly model the HVS and account for the first mechanism to detect not easily visible distortions [57], 3) our approach captures the low level features that can be used for object detection accounting for the second mechanism to detect supra-threshold distortions [57], 4) the extracted low level features can influence visual perception based on earlier findings [35], 5) visual saliency is incorporated in our framework by using our new center-surround processing step.

6.2 Computational Model

Our model contains four major components: Filter-bank decomposition, Feature distribution normalization, Spatial frequency scaling and Neighborhood pooling. The evaluation scores of two images can then be compared to assess their perceptual similarity or difference.

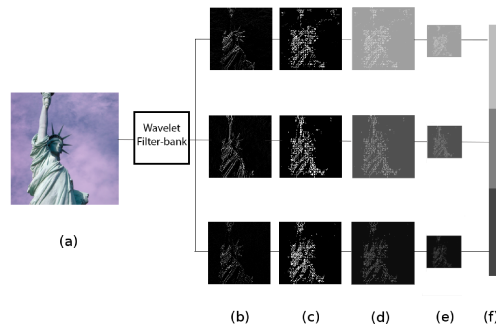


Figure 6.1: Outputs at various stages (a) The original image, (b) Three filter outputs, (c) Results after blockwise normalization, (d) Schematic representation of frequency scaling, (e) Pooling, (f) Pictorial representation of the final feature.

Pre-processing and filter-bank decomposition

An image is first converted into the CIELab color space generating the luma (L) and chroma (a and b) components. Each of the Lab components is then decomposed by a seven level wavelet, mimicking the frequency characteristics of the HVS [25]. We use BIOR 1.5 wavelets for this decomposition (Fig. 6.1 (b)). The output of these wavelet filters are represented by feature maps.

Note that we use a custom wavelet decomposition, rather than a trained set of filter weights described in the original model [42], the performance of which can be limited by the training set. A custom approach better simulates HVS characteristics like brightness induction and other frequency dependent ones [86].

Two-tier feature distribution normalization

The saliency of an image is dictated by the viewer’s perceptual power to discriminate between the center and surround along with the relative distribution of features in the target region [12]. Motivated by this “center-surround” processing in biological vision, a two-tier normalization process is incorporated in our model. We perform patch-wise divisive normalization and subtractive normalization after wavelet decomposition. A feature map is divided into patches of 13 x 13 with an overlap of 4 pixels between patches. The 5 x 5 center region within the patch is then processed. We use the surround to simulate the retinal eccentricity, beyond which the vision is blurred. Earlier studies [86] showed that this patch definition better simulates brightness induction of HVS.

Tier 1: Individual feature map normalization

A temporary subtractive normalization is first performed by subtracting the mean of the coefficients as shown in Eq. 6.1. $C_{s,o}(i, j)$ denotes the feature map at level s and orientation o of the wavelet decomposition for all pixels (i, j) surrounding the current pixel in the 5 x 5 center block, and the mean of the coefficients is: $\overline{C_{s,o}(i, j)}$.

$$v_{s,o}(i, j) = C_{s,o}(i, j) - \overline{C_{s,o}(i, j)} \quad (6.1)$$

Let $\sigma_{v_{center}}$ be the standard deviation of the feature map values in a 5 x 5 neighborhood around the current pixel (i, j) and $\sigma_{v_{surrounding}}$ be the standard deviation of the feature map values in the corresponding 13 x 13 neighborhood. We calculate the normalization factor r for the divisive normalization as:

$$r = \begin{cases} \frac{\sigma_{v_{center}}}{\sigma_{v_{surrounding}}} & \text{if } \sigma_{v_{surrounding}} \neq 0 \\ \sigma_{v_{center}} & \text{if } \sigma_{v_{surrounding}} = 0 \end{cases} \quad (6.2)$$

Divisive normalization is essentially a decorrelation performed by dividing each pixel $v_{s,o}(i, j)$ of the 5 x 5 neighborhood by r .

$$y_{s,o}(i, j) = \frac{v_{s,o}(i, j)}{r} \quad (6.3)$$

The mean values are then added back to each feature map pixel $y_{s,o}(i, j)$. The temporary mean subtraction enhances only the variation in the visual information, without altering the mean value of the feature maps.

$$V'_{s,o}(i, j) = y_{s,o}(i, j) + \overline{C_{s,o}(i, j)} \quad (6.4)$$

Tier 2: Cross feature map normalization

After $V'_{s,o}(i, j)$ is computed for each feature map, the mean value $\overline{V'_{s,o}}$ across all feature maps is calculated. A subtractive normalization using this average is then performed across levels, i.e., level $s \in (2, 7)$ and orientation o . $C'_{s,o}$ are the feature map values after the subtractive normalization.

$$C'_{s,o} = V'_{s,o} - \overline{V'_{s,o}} \quad (6.5)$$

The approximation feature map, i.e., level 1 (coarsest), is left unaltered in order to preserve the low frequency components that might have been lost on the final feature representation during the normalization and scaling processes.

Using this center-surround processing operation, we are able to simulate effects similar to lateral inhibition in the HVS, thus enhancing the regions of the image that have more variations. This has the added benefit of enhancing the visually salient regions of the image; e.g., global saliency computation in [40].

Liu and Heynderickx [68] improved the performance over SSIM [108] by scaling with the saliency term, following which they performed dissimilarity computation. We achieve the same effect in our method implicitly, by enhancing salient feature map values by our center-surround processing, before calculating perceptual distance.

Spatial frequency scaling

The normalization step described in the last section allows us to capture the local features within each level (feature map) of the image. However, the HVS

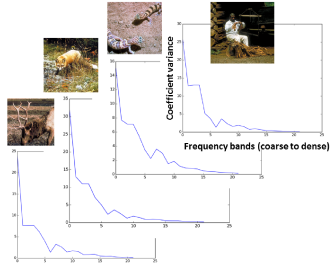


Figure 6.2: Example images illustrate the general trend of the coefficient variance associated with a feature map. It decreases when moving across levels.

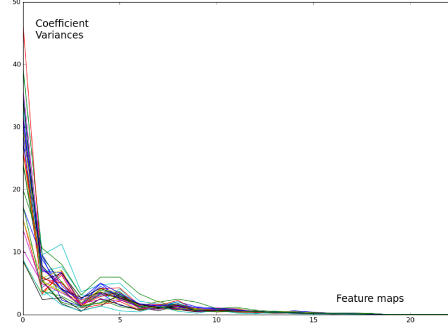


Figure 6.3: The consolidated result from thirty images also show that the variance coefficients in feature maps decrease when moving across levels.

has different sensitivities to various levels and orientation of image edges. We observe that as the feature map detail increases (from coarse to fine) across levels, the corresponding variance associated with the feature maps decreases (Fig. 6.2). We examined the trends in thirty images and they display this characteristic collectively (Fig. 6.3). Feature maps at finer (higher) levels contain more detail and have smaller standard deviations. Motivated by this observation, we introduce a new spatial frequency scaling formulation to project the local frequencies at different levels to a global space.

Feature map scaling

We control the projection using the standard deviation of the feature map at each level and orientation, i.e., $\delta(s, o)$. We scale each of the feature maps by $\delta(s, o)$. The frequency scaled coefficients ($C_\delta(x, y)$) are:

$$C_{\delta(s,o)} = \delta(s, o) * C'_{s,o} \quad (6.6)$$

where the scale factor is:

$$\delta(s, o) = K_2 / \sigma_{C_{s,o}} + K_1 \quad (6.7)$$

where s is level of feature map; o orientation of the filter associated with horizontal, vertical and diagonal feature maps after wavelet decomposition; and $\sigma_{C_{s,o}}$ the standard deviation of the processed feature map at level s and orientation o .

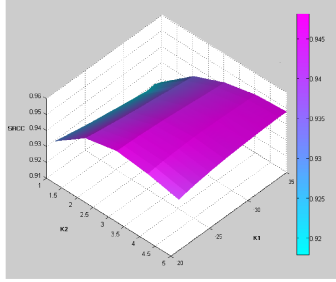


Figure 6.4: An optimization formulation can be obtained by adjusting the values of K_1 and K_2 .

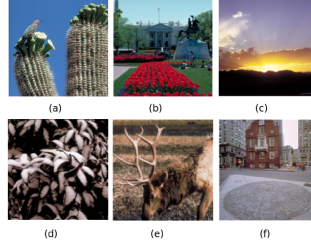


Figure 6.5: Colorful images (a),(b) and (c) are recognized by $Cr \geq 0.25$. Example images with $Cr < 0.25$ are shown in (d),(e) and (f).

An optimization can be formulated by choosing the appropriate values of K_1 and K_2 . For illustration, we plot the SSRC correlation scores in Fig. 6.4 using different pairs of K_1 and K_2 values. More detail about the correlation evaluation is presented in the results Section. The plot is the average result collected from 800 images. In our experiments, we set K_1 and K_2 to 31 and 3 respectively, which produced the optimal results. Coefficients of the frequency band and orientation that have higher scaling value $\delta(s, o)$ have a greater value in the final feature representation, corresponding to higher sensitivity to the HVS.

6.2.1 Color adaptation

It is believed that sensitivity of the HVS to scene content is derived from various stimulus modalities, including intensity, color, spatial and temporal (for dynamic scene only) features. This is consistent with what we noticed from our datasets, where brightly colored images display a different perceptual sensitivity compared to the less vibrant images. To address this issue, we assess the chroma component of an image. If the color ratio $Cr > 0.25$ (Eq. 6.8), we adaptively modify the frequency scaling function of each frequency band by $(K_2/\sigma_{L_s} + K_2/\sigma_{C_{as,bs}} + K_1)$, where $\sigma_{C_{as}}$ and $\sigma_{C_{bs}}$ are the standard deviations of the processed chroma feature maps at level s . K_1 acts as a decorrelation term for the color components of the image.

$$Cr = \sum_{s=1}^n \frac{\sigma_{as} + \sigma_{bs}}{(\sigma_{Ls})} \quad (6.8)$$

σ_{Ls} denotes the standard deviation of the feature map of the luminance component at level s . $\sigma_{C_{as,bs}}$ is the product of σ_{as} and σ_{bs} . The experimentally determined threshold of 0.25 was chosen to distinguish between colorful and normal images. A comparison of the two categories of images is shown in Fig. 6.5.

Pooling

Similar to object detection based models, the pooling component takes an $N \times M$ block in a processed feature map, divides it into $k \times k$ blocks and returns a single value equal to the maximum valued coefficient in the block, helping us to achieve invariance to small translations. We choose k as 3 for a window size of 3 x 3 and select the maximum of the feature map values in the block. Assuming the dimensionally reduced feature map as $C_f(s, o)$, the final feature representation f for an image, we apply concatenation on all the processed feature maps $C_{f(s,o)}$.

$$F = concatenate(C_{f(s,o)}) \quad (6.9)$$

for every level s and orientation o of the processed feature map $C_{f(s,o)}$.

Perceptual distance measurement

In order to compute the perceptual similarity between images, we use the $L1$ norm to compare the features between images.

$$e = \sum_{i=1}^N |F_1(i) - F_2(i)| \quad (6.10)$$

F_1 and F_2 are the two features generated from Images 1 and 2 respectively; N is the dimension of the features; and e is the perceptual difference of Image 1 with reference to Image 2.

		Proposed	VSI [125]	PSNR	SSIM[108]	MS-SSIM[111]	VSNR [14]	VIF[98]	FSIM [127]	IW-SSIM [97]	IFS [15]	GSIM [67]	MAD[57]	GMSD [1]
LIVE (2005)	SRCC	0.9430	0.9524	0.8756	0.9479	0.9513	0.9274	0.9636	0.9634	0.9567	0.9599	0.9554	0.9669	0.9600
	KRCC	0.8200	0.8058	0.6865	0.7963	0.8045	0.7616	0.8282	0.8337	0.8175	0.8254	0.8131	0.8421	-
	PLCC	0.9467	0.9482	0.8723	0.9449	0.9489	0.9231	0.9604	0.9597	0.9522	0.9586	0.9437	0.9674	0.9600
TID (2008)	SRCC	0.8820	0.8979	0.5531	0.7749	0.8542	0.7046	0.7491	0.8805	0.8559	0.8903	0.8554	0.8340	0.8910
	KRCC	0.6969	0.7123	0.4027	0.5768	0.6568	0.5340	0.5860	0.6946	0.6636	0.7009	0.6651	0.6445	-
	PLCC	0.8883	0.8762	0.5734	0.7732	0.8451	0.6820	0.8084	0.8738	0.8579	0.8810	0.8462	0.8306	0.8710
CSIQ (2010)	SRCC	0.9432	0.6423	0.8057	0.8755	0.9132	0.8105	0.9194	0.9242	0.9212	0.9581	0.9126	0.9467	0.9560
	KRCC	0.7879	0.7857	0.6078	0.6900	0.7386	0.6241	0.7532	0.7561	0.7522	0.8158	0.7403	0.797	-
	PLCC	0.9395	0.9279	0.8000	0.8612	0.8991	0.8002	0.9278	0.9120	0.9144	0.9576	0.8979	0.9502	0.9540
TID2013	SRCC	0.8829	0.8965	0.6394	0.6274	0.7851	0.6818	0.6769	0.8015	0.7779	0.8697	0.7846	0.7808	-
	KRCC	0.6979	0.7183	0.4696	0.4554	0.6029	0.5084	0.5147	0.6289	0.5977	0.6785	0.6255	0.6035	-
	PLCC	0.8890	0.9000	0.7017	0.6861	0.8334	0.7129	0.7720	0.8589	0.8319	0.8791	0.8267	0.8267	-

Table 6.1: Comparison of performance on datasets.

6.3 Experimental results

We tested our framework on four benchmark databases, which contain a set of original images, the degraded version and the perceptual quality scores, i.e, mean opinion scores (MOS). We generated quality scores on these images using our algorithm. For a fair comparison with other algorithms a logistic function was fitted to get a non-linear mapping from the objective scores to the subjective scores, following [95]. The comparison was based on Spearman rank order correlation coefficient (SRCC), Kendall rank-order correlation coefficient (KRCC) and Pearson linear correlation coefficient (PLCC). A good IQA is characterized by higher values for SRCC, KRCC and PLCC. Our implementation in python had an average time per image pair (1 reference and 1 distorted) of 3.06s for the LIVE dataset. This can be improved with C++ programming. Our computational complexity is similar to that of SSIM; one wavelet decomposition (FIR implementation $O(N \log(N))$) followed by a scaling in windows and then on sub-band level. Since a wavelet transform is used, the total number of pixels in all the sub-bands remains the same as the original image, resulting in low overall complexity.

6.3.1 Performance comparison and analysis

The strengths of our model can be attributed to the local normalization and global scaling processes, which have advantages over traditional methods. These operations model the adaptation of the HVS by decorrelating elements in the feature maps along the axis of the wavelet basis, mimicking the processing in retinal ganglion cells [8] [28]. The end result is similar to PCA whitening (where the whitening operation makes the different components of PCA un-

correlated and of unit variance). The color intensity invariance feature of our model improves the IQA results by adaptively categorizing images at different brightness levels and analysing accordingly.

To test the performance, we run our IQA algorithm on the CSIQ [58], LIVE [37], TID2013 [89] and TID [91] databases. A comparison among different IQA algorithms on various datasets is given in [116], which we use to evaluate our algorithm. The algorithms that we compare with are VSI (Visual Saliency induced IQA) [125], PSNR, SSIM (Structural similarity based IQA) [108], MS-SSIM (Multi scale SSIM) [111], VSNR [14], VIF (Visual information fidelity based IQA) [98], FSIM (Feature similarity index IQA) [127], IW-SSIM (information weighted SSIM) [97], IFS (Independent feature detector IQA) [15], GSIM (Low level gradient similarity IQA) [67], MAD (Most Apparent Distortion) [57], GMSD (Advanced SSIM based on gradient magnitude IQA) [1]. The results are shown in Table 6.1.

Over the years, many IQA algorithms have been introduced. In order to evaluate the correlation between subjective scores and objective scores generated by IQA algorithms, statistical ranking methods like SRCC, KRCC and PLCC are used. However, what is the significant threshold in these rankings which truly reflects noticeable visual quality difference in the assessed images? Is it 0.01 or 0.05? In the VSI paper, the authors highlight the top two scores with a difference up to 0.05. In Table 1, we bolded the scores which are within 0.03 (half way between 0.01 and 0.05) of the maximum value. Our method has all bolded scores while others have at least one score not bolded. For comparison, if the threshold is reduced to 0.02, the proposed method only has one not bolded score while IFS and VSI have 3 and 6 respectively. One obvious reason for our consistent performance, as illustrated in Table 1, is that other algorithms work particularly well in one test database at the expense of another. For example, LIVE was released around 2005 by the authors/co-authors of [SSIM, MS-SSIM, VIF, FSIM, IW-SSIM and GMSD] (published in 2004, 2003, 2005, 20011, 2009 and 2013 respectively). The techniques described in these papers, following the concept of SSIM, all perform well in LIVE. Image structural content is an important factor for quality assessment, but an algo-

rithm designed for one type of structure, e.g. edges, may not be effective on another. Besides, we observed that by tuning the parameters in an algorithm, the outcome may favor one test dataset over another. By adopting an optimal set of parameter values, as in our method and as explained in the IFS paper, a balance in quality across datasets can be achieved.

Among the state-of-the-art IQA techniques from 2005 to 2015 no one algorithm performs best for all datasets. GMSD does not show KRCC score and the test on TID2013 is missing. Thus, we exclude it from the comparison. Both IW-SSIM and MAD were published in 2009. While IW-SSIM outperforms MAD in TID, MAD is better in CSIQ and LIVE. Both algorithms work equally well in TID2013. While MAD works better in all four datasets than VIF (2005), IW-SSIM is not as good as VIF in the LIVE dataset. FSIM was published in 2011. Although it shows improvement over IW-SSIM in the TID2008/2013 datasets, MAD (2009) is better than FSIM in the CSIQ and LIVE datasets. VSI and IFS were published in 2014 and 2015 respectively. VSI shows the best results in the two TID datasets and IFS is better in CSIQ, but MAD is still the best in the LIVE dataset.

Since IQA research has advanced rapidly in recent years, new parameters have been introduced in the algorithms to accurately assess image content. Accordingly, small image datasets are expanded to increase the variety of image content. It can be seen that compared with TID2008, TID2013 has seven extra types of distortions adding up to a total of 24 types. In comparison, LIVE (Release 2) has only five distortion types. Since TID2008 can be treated as a subset of TID2013, we exclude the scores of TID2008 to avoid double counting. Also, as pointed out in the IFS paper, “independent component analysis can provide a good description for the receptive fields of neurons in the primary visual cortex which is the most important part of the HVS.” Image contents vary and each image is composed of low level components which stimulate the HVS. SSIM-based techniques detect certain types of component successfully, e.g., edge structures in the LIVE dataset. However, there are other perceptual components, such as luminance and color, generated from different types of distortion which are not described in the LIVE dataset. Thus, evaluation

based on the scores in LIVE does not truly reflect potential distortions. Since TID2013 contains 3000 images and CSIQ (2010) contains 866 images, which are far more than other datasets, and they were created more recently, we use them as benchmark datasets for evaluating IQA algorithms.

In a real-world application, it is not possible to predict what type or what combination of distortion(s) will occur in the processing, transmission and rendering pipeline, and accordingly select the best performing algorithm. Instead of comparing scores for 24 distortion types individually, it is practical to examine the overall (average) performance of an algorithm. Based on the test datasets CSIQ and TID2013, the 5 best performing algorithms with average score over 0.8 are MAD, VSI, IFS, FSIM and our proposed method.

Note that IFS and VSI have the best performance in the CSIQ and TID2013 datasets respectively, but at the expense of the other dataset. FSIM, IFS and MAD have a high difference in average score of more than 0.1 between the two datasets. Our algorithm has a high average and achieves more consistent performance in both datasets. Our method shows an improvement compared to low level feature based IQA FSIM [127]. The inclusion of two-tier normalization, optimized frequency scaling and color adaptation, attributes to this improvement.

To-date, there has been no one single IQA algorithm which outperforms others for all distortion types and in all benchmark test datasets. Our contribution lies in proposing a more consistent technique to assess image quality based on a systematic approach to review the evolution of IQA algorithms using unbiased test data, instead of following the traditional method to look at individual scores in isolation. The scatter plots of the scores generated by our metric against user subjective scores are shown in Fig. 6.6, which illustrates the consistency between our metric scores and the user scores. In order to illustrate the advantage of using center-surround in the normalization process, we computed the scores using only a simple window of size 3x3, 5x5 and 7x7 respectively. The results are shown in Table 6.2. The lower performance of single window compared with center-surround agrees with the finding that the saliency of an image is influenced by the relative distribution of the image

		Center-Surround	3x3	5x5	7x7
CSIQ	SRCC	0.9432	0.9369	0.9346	0.9292
	KRCC	0.7879	0.7792	0.7752	0.7662
	PLCC	0.9395	0.9212	0.9183	0.9124
	RMSE	0.0902	0.1026	0.1044	0.1079

Table 6.2: Performance of using a single window compared with using a center-surround analysis.

features in the center and surround [12].

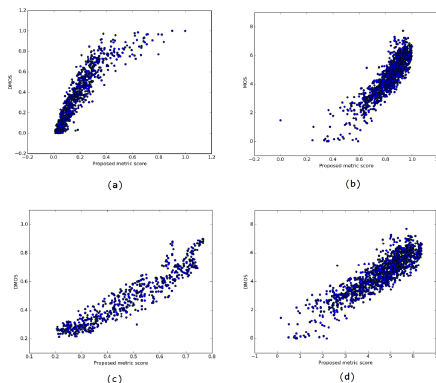


Figure 6.6: Scatter plots of our metric scores vs subjective scores on dataset (a) CSIQ, (b) TID, (c) LIVE and (d) TID2013.

It is worth pointing out that MAD performs well in CSIQ but not in TID. IGM performs well in TID but not in CSIQ. Our proposed method has more consistent performance in both databases. This advantage is particularly clear in our RMSE scores in both datasets.

An IQA algorithm often works well in one dataset but not in another. This is similar to machine learning algorithms whose performance depend on the training set. Depending on how the controlling parameters of an IQA algorithm are tuned, e.g., tuned with the CSIQ dataset, test images sharing similar characteristics with the training dataset will perform better using that algorithm. In order to illustrate this point, we fine-tuned the parameters K_1 and K_2 using the CSIQ dataset only. The optimal values were 33 and 5 respectively, leading to an increase of SSRC from 0.943 to 0.948 (Fig. 6.4), which is higher than the value 0.946 of MAD. However, the fine-tuned parameter values lead to a decrease of SSRC from 0.882 to 0.869 in the TID dataset. Until there is an IQA algorithm which can adaptively adjust the controlling

parameters at the image level, it is difficult to have one algorithm that outperforms others in all image datasets. Given an arbitrary image, without knowing what training characteristics it is associated with, our method guarantees a balanced assessment.

Chapter 7

Investigation of Gaze Patterns in a case study

7.1 Introduction

One practical application of our results on low-level features and attention is in the field of surgical training. Specifically, we focus on Laparoscopic surgical training. Laparoscopic Surgery (LS) is becoming the standard procedure in surgery. It uses small incisions created in the patient's body to insert surgical tools and camera. The operation is performed with the surgeon guiding the surgical tool with the camera guiding his movements. LS offers a lot of benefits to the patients including reduced recovery times, less risk of hemorrhaging etc. However, for the surgeons using the system, it is difficult to perceive the surgical site in 3-dimensional (3D) fashion and coordinate their eyes and hands, due to the loss of depth perception, indirect image, mirrored hand movements, and eye-hand misorientation using a single camera[11]. The limitation of visual perception in LS increases cognitive and physical stress of the surgeons and trainees and is a leading cause of inaccurate judgment and estimation. This leads to significantly longer times in training and performing LS **c2**

Some recent studies have found that different camera arrangements affect perceptual-motor performance in laparoscopic surgery [31] - [38]. The use of multiple cameras as a tool for restoring the three dimensionality is optimistic and can easily resemble the different vantage points accessibility of open surgery. In the current study we investigate the behavior of subjects

when presented with multiple viewing perspectives in surgical simulation. We compare the eye behavior of the high and the low performers when attempting to perceive the depth cues presented with a multiple view setting.

7.2 Motivation

It was shown that a multiple view arrangement can be superior to the use of a single camera [31] - [38]. However, this also increases the cognitive load on the user. From studies in aviation displays, we have the conclusion that mentally integrating information across multiple displays is challenging and draws additional attention demands from the user [84] - [115]. In the study of DeLucia on effects of camera arrangement on perceptual-motor performance in LS, multiple-camera views provided more information about 3D space but imposed more attention demands compared with a single-camera view. In their study participants were presented with multiple-camera views of a surgical simulation environment. Participants did not look at all views equally often and may not have necessarily mentally integrated the views to reconstruct 3D space [22]. It was also suggested that surgeons (elite performers) use different information or integrate multiple sources of information differently than novices. This leads us to believe that with training, humans might learn a specific 'gaze behavior' that integrates the 3D information more efficiently. We seek to discover this behavior that separates the experts from the novices. The results can help in design of better displays in multi-view environment, more intelligent camera placement and better training programs for novices [107].

7.3 Method

In the study we conducted, we compared the gaze behaviors of human operators while performing simple surgical task in a multiple camera view condition.

7.3.1 Subjects and experimental environment

Twenty university students with varying levels of laparoscopic surgical training participated in the study (13 males, 7 females; mean age, 28). The sample size was calculated based on outcomes from previous research found in the literature. For example, we used DeLucia's study (2011) to adjust our sample size. In DeLucia's study, 12 subjects were included in testing how the number and type of camera views affect manual manipulation. They recorded a main effect of viewing condition ($F(4, 44) = 23.79, p < 0.001, \eta^2 = 45.58\%$). Tukey's HSD analyses showed that mean task completion time was significantly faster for the direct view compared with the front and side views, and the side view resulted in the slowest completion time among the different views ($p < 0.05$). The larger group size in the current study was expected to have significant power to show significant effects. Ethics approval was obtained from the Health Research Ethics Board of the University of Alberta before the recruitment of human subjects. Written consent was obtained from each participant prior to entering the study.

The experimental setup included three main components, (1) a 2D monitor (LG-24MA31D, LG Electronics, Seoul, South Korea) which displays images captured by the surgical cameras (2) Training box with three camera for front, top and side views and 3) Tobii X2(Tobii Technology, Inc., Washington DC, USA) 60Hz eye-tracker placed under the monitor to unobtrusively record the subject's eye motions.

Subjects performed a surgical simulation task with the camera placed at two different angles (front camera at 30-, and top camera at 90- degree angle to the plane of the target), with a third camera placed on the side of the training box.

7.3.2 Tasks

Subjects were asked to move graspers and transfer objects between surgical targets (pegs) in different depth planes (Figure 3). Subjects were required to perform the task as quickly and accurately as possible without dropping

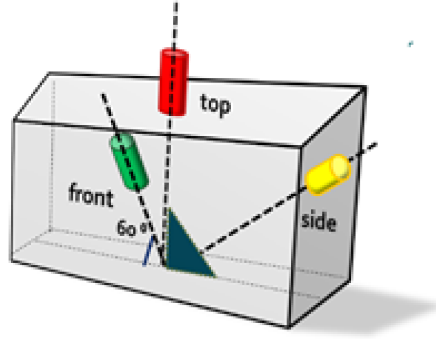


Figure 7.1: 2D camera positions in training box.

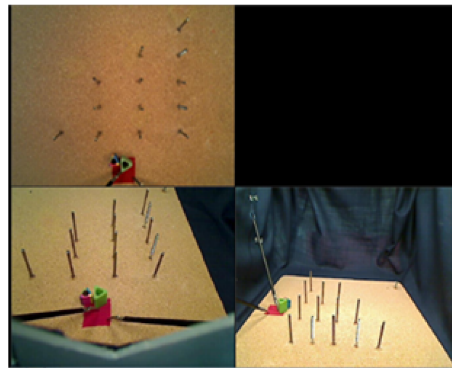


Figure 7.2: Three views.

the objects. During the entire trial subjects' eye movements were remotely recorded by an eye-tracking system

7.4 Analysis

Careful analysis on eye-gaze behaviors of the higher and the lower performance trials was conducted by analyzing the eye tracking data from Tobii Studio. We also built a proof of concept system for analyzing the object distances and velocities at each subtask. To do this, we used Kanade-Lucas-Tomasi (KLT), feature-tracking algorithm [10] for tracking the objects. The algorithm can be used to track a set of feature points using optical flow estimation. We used an opencv implementation of the same.

Our implementation re-initializes the tracking every 10 frames or every time there is less than 2 feature points that are tracked. The feature points

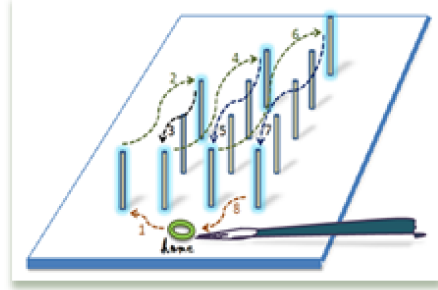


Figure 7.3: Task that the subjects were asked to complete.

we tracked are found using color thresholding and feature detection algorithm proposed in [100] that detects strong corners of the image from the video frame. A screen grab from the actual system showing the object detection and tracking is shown in Fig 7.4.

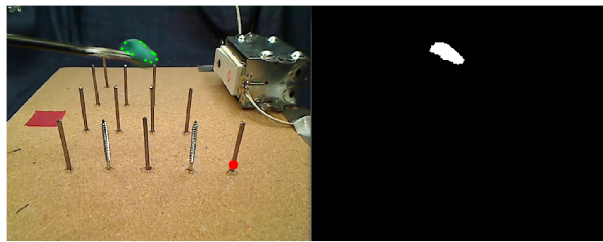


Figure 7.4: Object detection and tracking; the green spots denote the tracked points, the red spot denotes the target point. The figure on the right shows a segmented view of the pixels being tracked.

Detection of task completion and next target: In the series of tasks we asked the user to perform, the target kept changing after the user performed the task. We asked the experimenter to mark all the targets of the task that the user has to perform. The system would cycle through these points to detect task completion.

7.4.1 Identification of experts using time performance

We used time of task completion metric to measure performance. All 20 subjects performed the study task in the multiple viewing condition. The average completion time was 4.89 min (min: 2.2; max: 8.37). We selected the fastest 25 % as high performers and the slowest 25% as low performers for conducting further analysis.

Statistical tests: By comparison of the low and high performers' eye behavior we were able to reveal how human operator collect visual information to rebuild 3D vision in Image-guided surgery. The eye behavior measurements (percentage of view used and frequency of gaze shift) were subject to 2 (groups; high performers versus low performers) x 2 (views; top versus front) two-way ANOVA model for analysis of variance. To define further difference between the high and the low performance groups, data was subjected to a independent samples t-test. SPSS (SPSS Inc. Chicago) was used to perform statistical analysis and $p < 0.05$ was considered statistically significant.

7.4.2 Features used

From the eye tracker and the feature tracking, we analyzed the following data:

1. Gaze location: Initial analysis of subjects' gaze location over the three camera views subjects spent sufficient time (48%) on the top view as well as the front view (50%), but not on the side view, which was used at only 1%. Additionally, we compared the number of visits between the 3 views. Pair comparison revealed significant difference between the top and the side view ($p < 0.001$); front and side view ($p < 0.001$), but not between top and front view ($p = 0.971$). For this reason, we excluded the side view data in our further comparisons.
2. Frequency of gaze shift between the different views: This looked at the amount of time there was a shift of gaze from one view to other. The statistical test results showed that there is significant main effect between the high and low performers : $F(3,16) = 8.96$; $p = 0.009$; $\eta^2 = 0.359$. The graphical representation of the same is show in 7.5
3. Percentage of view used (eye behavior measurements): This looked at the amount of time one particular view was utilized during the course of the experiment by a user. We consider percentage values here to remove the effect of varying times among users. Results revealed for the group as a whole, significant main effect for views: $F(3,16) = 27.71$; $p = <$

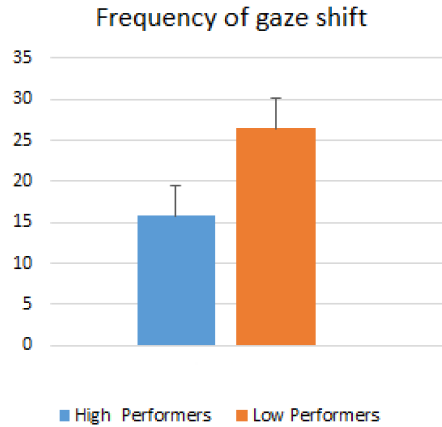


Figure 7.5: Number of times the Gaze of the user shifted between top and front view by the high and low performers.

0.001; $\eta^2 = 0.634$; but not significant main effect for groups: $F(3,16) = 0.02$; $p = 0.89$; $\eta^2 = 0.001$; and significant interaction effects: $F(3,16) = 6.56$; $p = 0.02$; $\eta^2 = 0.29$. Results are in Fig 7.6

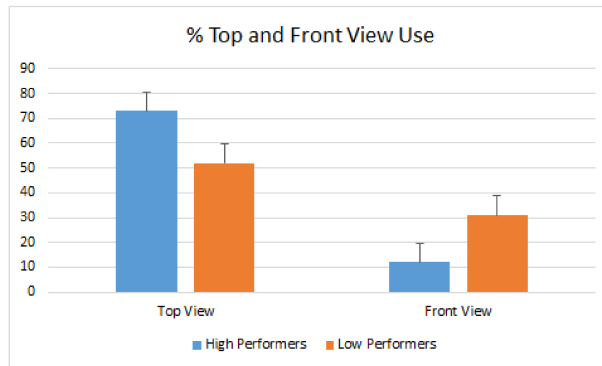


Figure 7.6: Percentage of Top and Front view used by the high and the low performers.

4. Front/Top ratio is obtained by dividing the percentage of gaze on the front view to the percentage of gaze on the top view. The Front/Top ratio was significantly different between high performers ($M = 0.19 \pm 0.15$) and low performers ($M = 1.02 \pm 1.12$) groups: $t(8) = -1.64$; $p = 0.034$. Graphical results are shown in Fig 7.7

5. On screen distance (δ) defined as:

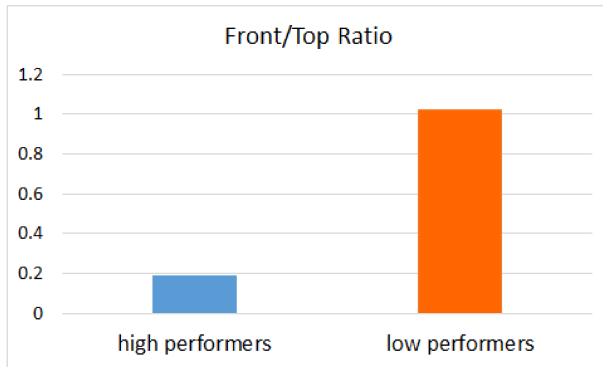


Figure 7.7: Front/Top ratio for high and low group performers.

$$\delta = \min(d_{top}, d_{front}) \quad (7.1)$$

where d_{top} , d_{front} are the on screen distance of the tracked object to the target in the top and front views respectively. A plot of on screen distance vs amount of time front and top views were used for the experts and novices for the use of top view and the front view are shown below for distances up to 200 pixels (Sample plots of high and low performers in Fig 7.8, 7.9).

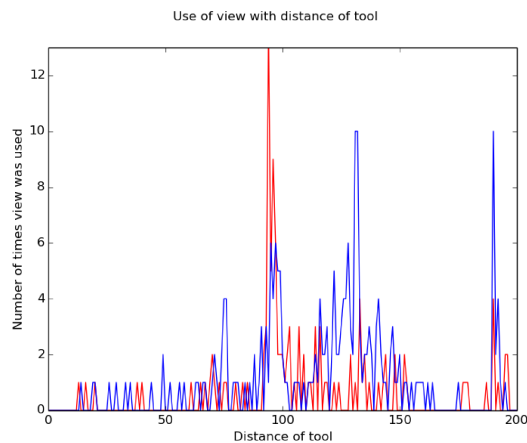


Figure 7.8: High performer. Here red plots are the cases where the top view was used. The blue are the cases where the front view was used.

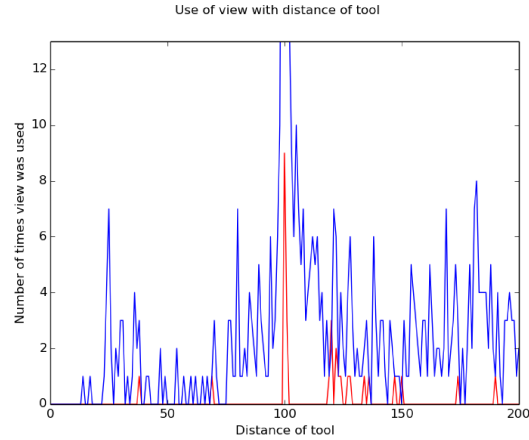


Figure 7.9: Low performer. Here red plots are the cases where the top view was used. The blue are the cases where the front view was used.

7.5 Conclusion

Based on the results of our study, we can underline a few key points of importance. We confirmed the theory that additional views contain important information for subjects to perceive the depth of the surgical site. This was proven with the finding that 48% of the time, subjects used the top view for guiding their performance along with the conventional (front) view used at 51%.

Our results revealed that the high performance group used the top view more than 70% of the trial time compared to the front view, which they referred to only 12% of the time. This was particularly so when the object distance was closer to the target. This behavior was confirmed by the detailed analysis of the camera use at specific distances from the target shown in Fig 7.8,7.9. Other time was spent on view transition, saccades, blinking, and side view visits.

We found that the members of the high performance group shifted their gaze between the top and the front view less frequently compared to their low performance counterparts. This finding suggests that the “experts” performers exhibited more concentrated to a particular view eye behavior whereas the “novices” were frequently shifting their visual attention between views

High performers used both views in more balanced way throughout the

trial compared to their low performance counterparts which focused mainly on one view. We conclude these from the Front/Top ratio for the high and low performers.

In conclusion, the current study showed that human operators utilize information from different visual sources, when available, for reconstructing the three-dimensionality of a surgical scene without impairment of performance. We additionally revealed eye behavioral evidence to support the notion that expert and novice performers use the visual information from the multiple sources differently. As was suggested in previous research on perceptual-motor performance in the LS [31], top surgeons (elite performers) might use different information or integrate multiple sources of information differently than novices. It is also believed that, with practice, trainees could learn to use the multiple views to improve performance beyond what was achieved with a single view. Thorough understanding of the operators' behaviors is vital to direct further research on the feasibility of multiple views usage in the OR and for surgical training. The knowledge from this study can be used for creation of smart display interfaces where trainee's gaze can be guided towards an expert-like behavior. The benefit of such a remote training tool (smart display system) is significant.

This study was useful in analysis of expertise in a task specific environment. A similar experiment can be performed to analyze the effect of human experience with art on perceptual quality.

Here there is a big potential that machine learning can be used to capture and model differences in gaze behavior. The results of our studies can be easily modeled by a simple system like an SVM. since we derive a set of features that can be separated by a linear hyperplane. More complex behavior can be captured by deep learning based systems that work on the raw data. A potential future system can be build that can monitor gaze behavior to check for expertise level per session. The analysis can also be done in time windows hence having a fatigue tracking mechanism.

Chapter 8

Impact of JPEG compression on perceptual quality of 3D models

8.1 Introduction

3D graphics is an integral part of modern multimedia with applications in various areas including video games, medical data transmission, and virtual reality. 3D models are a fundamental part of 3D computer graphics. The most common method to represent 3D model is by using a triangular mesh to describe 3D geometry and one or more 2D images to represent the texture. We will be referring to 3D models used in graphics represented this way as *tex-mesh* in this paper.

With the recent push towards cloud storage and interactive 3D graphics, a significant part of the information exchange is happening over the Internet. The information in 3D graphics consists mainly of tex-meshes. Given limited bandwidth and a high amount of tex-mesh data to be transmitted, there is a significant drop in transfer speed and hence the quality of user interaction with a graphics application can become unacceptably slow. One way to increase the speed is to compress the data, which leads to loss in perceived quality. This in turn leads to a bad quality of experience for the user. Thus, it is extremely important to find a tradeoff between compression and perceived quality. Such a trade-off requires a model that can predict the perceived quality of a tex-mesh at a given compression level.

A popular method for increasing the efficiency in the representation of

multimedia data, that is meant to be perceived by the human visual system (HVS), is to remove perceptually irrelevant information. Identification of perceptually irrelevant information requires subjective studies, where user opinion about quality is collected. Since this is not feasible in every situation, a perceptual quality model can be used to mimic the HVS. Perceptual quality models analyze a given tex-mesh and give a score or rating to the quality of the tex-mesh that would be similar to what a human would score or rate.

The literature is rich with studies that model the perceptual quality of 3D meshes (without texture). However, the texture is an important part of the tex-mesh representation, as it contributes to the visual appeal of a model as well as act as a mask to hide some geometric imperfections in the mesh. Despite its importance, there is insufficient amount of research that deals with perceptual effects of texture on tex-mesh. We found no study that specifically deals specifically with texture compression and its effects on perceived quality of tex-mesh.

8.2 Experimental setup

Perceptual experiments were conducted with a rating based subjective experiment on a custom interface. The design and details of the interface are detailed in this section.

8.2.1 Design consideration

Number of reference: Providing a reference to subjects while taking opinion scores allows precise control for the variable we want to study (in this case texture compression). However, in studies like these, the self consistency of results is extremely important. The importance and positive results of these are shown in [87] and analysis is found in [30]. To ensure the consistency of results, we provide the subject with the best and the worst models as reference.

8.2.2 Interface

The interface developed for the study showed the users three different models. The leftmost model was the model with the maximum amount of texture compression (Q factor 1), the rightmost was the best texture quality possible (Q factor 100) and the middle was a model with variable texture quality. The user could interact with the model by using the mouse. These interactions were synchronized for all three models displayed. The interactions allowed were rotation and scaling of the models. The ratings were set on a The interface allowed recording of the user opinions by clicks on ratings labeled with adjectives: "Bad", "Poor," "Fair," "Good," and "Excellent.". This translated to a rating score in range [0,5]. The interface is blanked after a time limit of 30 seconds. The user can rate during this period, but cannot see the model.



Figure 8.1: Screen shot of the interface used for rating based experiment.

Selection of next model Upon selection of a rating, the interface selects the next model in a pseudo-randomly. This goes against the conventional practice of a purely random selection. However we found that random selection gives highly skewed results with few models having very few user responses. To ensure that we have a uniform number of user responses for each quality we follow an alternate procedure. Till we obtain a minimum number of user response per model, we maintain a pool of models that do not have enough user responses. We then randomly select from the models in the pool and display it for a subjective rating. Once a model has enough number of responses, the model

is removed from the pool. Once we ensure that all the models have a certain minimum number of response, we select models whose response have larger variance by using the same model pool based method above. This ensures that the models where there is large variance in user score is examined by more number of users; hence reducing our error.

8.2.3 Experimental conditions

The subjective experiment was performed with the interface displayed on an AOC 27 inch screen with factory color calibration. The subjects viewed the screen at a distance of approximately 70cm. The subjects had 30 seconds to take a decision. The computer used to perform the experiment had an i7 processor with 16GB RAM and and Nvidia GTX660 GPU.

8.2.4 Stimuli generation

We select from a total of 6 models. The visualization of the models are shown in figure 8.2. The details of models used are detailed in table 8.1. We limit the resolution of the model to a width of 480 pixels.

Model	#Vertex	#Texture shape	File Size	
			Mesh	Texture
Apple	6738	480x480	1.04MB	184KB
Bison	3821	480x480	883KB	188KB
Dwarf	6167	480x480	1.12MB	188KB
House plant	6608	480x480	58KB	61KB
Treasure chest	6792	960x480	43KB	294KB
Barrell	42284	480x480	7.03MB	100KB

Table 8.1: Details of 3D models used for psychovisual experiment.

8.2.5 Experiment

Each subject was initially briefed about the purpose of the study and informed about the reference models, time limits and mode of interaction in a short



Figure 8.2: Visualization of models used in experiment.

demonstration session. The subjects were then asked to rate center model on the scale of the interface. Each user is asked to rate 30 times.

We chose from a set of 6 different models with 6 levels of texture compression (Q factors 1,20,40,60,80,100) giving us 36 textured models to assess subjectively.

8.3 Mathematical Modeling

An initial analysis of the data can be performed by analyzing the average user rating for each Q factor. The plots of raw data collected from 9 subjects are shown in the plots. Here we calculate the error bar assuming that the sampling distribution is Normal and the standard deviation is unknown. Here we see that even without any processing the data that we obtain an average score that is monotonically increasing with Q factor; consistent with general intuitions on effect of compressed texture on quality.

The interaction between texture and geometry on perceived quality was modeled by [87]. The model generalizes texture and geometry quality with normalized variables for texture and geometry level. Since texture resolution and the range of quality levels are normalized in a fixed range, we can model the overall effect of texture compression by modulating the texture variable t in model proposed above.

Outlier Detection

Detection and removal of outliers are performed by using modified z score method proposed in [39] because of its robustness. We follow the recommen-

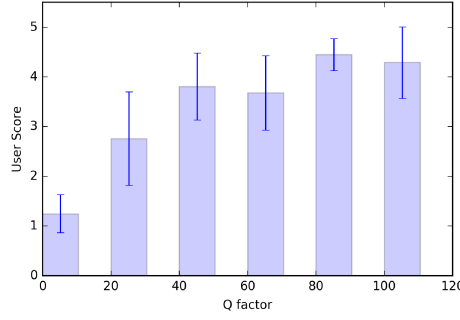


Figure 8.3: Average score of users for all models vs JPEG Q factors.

dations of the paper and remove the points above $M_i > 3.5$.

$$M_i = 0.6745 \frac{x_i - \tilde{x}}{MAD} \quad (8.1)$$

where x_i represents the i^{th} user score for any particular model, \tilde{x} represents the median of the user scores and MAD represents the median absolute deviation.

8.3.1 Analysis

After outliers removal, an analysis of the effect of compression on quality can be visualized by looking at the overall quality score for each of the compression level. On this data, we can fit a generic log function of the form $y = a * \log(b + cx)$ where x is the Q factor and a, b, c are the parameters of the fitting. The motivation for this stems from the fact that generally the quality drops observed in 2D image compression follow this trend. The results are shown in fig 8.4. Similar to observations of texture resolution in previous studies, here also, we find that till a certain compression level, we have extremely small drop in quality. Hence transmission of data with high Q factor leads to wasteful utilization of the bandwidth.

The integration of texture into mesh is modeled by equation 8.2 as per the results of [87].

$$Q(g, t) = \frac{1}{\frac{1}{m+(M-m)t} + (\frac{1}{m} - \frac{1}{m+(M-m)t})(1-g)^c} \quad (8.2)$$

Where m and M are the minimum and maximum bounds of quality, g and t are graphical and texture components scaled into a $[0,1]$ interval, and c is a

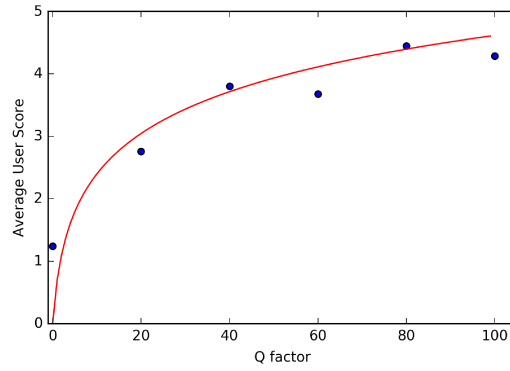


Figure 8.4: Change in perceived quality for each Q factor.

constant.

Our observations of texture compression can be integrated into this form by substituting t with \tilde{t} described by equation 8.3.

$$\tilde{t} = t * (a' \log(cx + b)) \quad (8.3)$$

$$a' = \frac{a}{5}$$

We can determine a, b, c by solving for these with the data we have from our experiment. We multiply a by $\frac{1}{5}$ to scale the data in same range of original experiment. From our results, we found the values of a, b, c is 1.02, 1.13 and 1.08 respectively.

Information about the file size of the texture also provide the same information albeit in a different form.

$$\tilde{t} = t * \frac{a'}{b - \exp(cx)} \quad (8.4)$$

$$a' = \frac{a}{5}$$

As stated earlier, in our experiment we standardize the texture dimensions for controlling the number of unknown variables in the experiment. Hence, we find that the overall texture sizes are generally similar (refer to table 8.1). An interesting observation here would be an analysis of how the image file sizes (which is actually impacting the bandwidth usage) affects quality. The scatter plot of the same is shown in fig 8.5. These observations match the observations

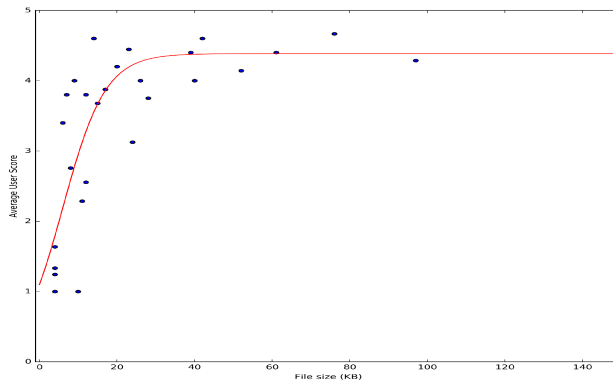


Figure 8.5: Change in perceived quality for different texture file sizes.

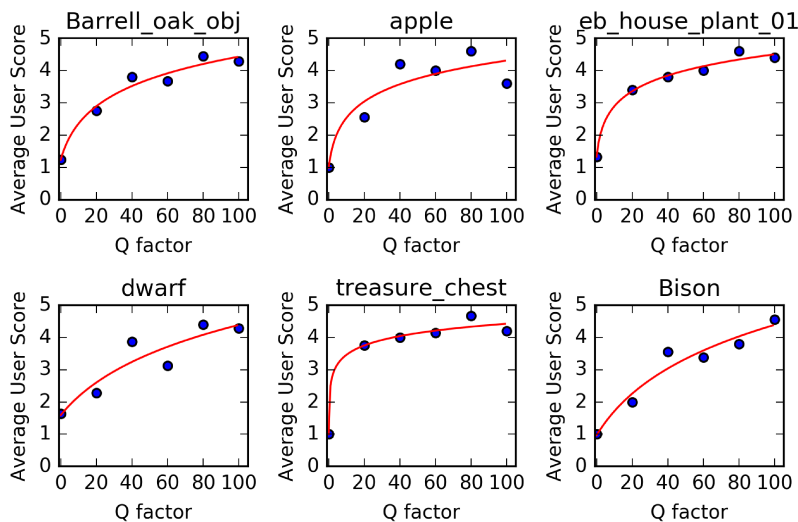


Figure 8.6: Average user score for each model at different levels of compression.

of mesh resolution [87]. Similar to the mesh quality, we can fit a curve of the form $y = \frac{a}{b - \exp(cx)}$ with parameters a, b, c . We see a saturation of perceived quality after a certain file size. The upper limit would be M as described in the original paper. From our results, we found the values of a, b, c is 1.08, 0.24 and 0.19 respectively.

8.3.2 Future work

Further insights can be obtained by fitting curves are fit individually on ratings of each of the models with different compression factors. The results of this

step is shown in fig 8.6. An interesting observation here is that certain models where the models are animate(alive), we find that the decrease in quality is more linear. This opens up the possibility of model content influencing the rate at which the global texture quality deteriorates. This however is beyond the scope of this work and can be considered as future work.

8.4 Conclusion

We studied the effects of JPEG texture compression on the perceived quality of a textured 3D model. We performed a rating based subjective experiment to measure human responses to degradation in texture because of JPEG compression. We then model our observations by statistical model and integrate it into existing model by [87]. We also provide an alternative model based on file size and integrate the same into this model also.

Chapter 9

Conclusions and future directions

In this work, we proposed various methods for expanding the field of image quality assessment.

Our major contribution is the proposal of the first no reference image quality assessment method for evaluation of local image quality in High Dynamic range images. To address the issue of algorithms failing to predict the visual quality in HDR images, we derive a neural network based solution for estimating errors and deriving the perceptual resistance in an image. Perceptual resistance is a measure that we define to denote the degree of difficulty of a viewer to perceive an error in a given part of the image. We derive this measure using a purely data driven method on IQA datasets without the need for psychovisual experiments(which is the traditional method to derive this measure). Further experiments verified the performance of perceptual resistance with experimental evidence on visibility on LDR images. We then studied the impact of texture compression on 3D textured models and modeled our findings into existing models for perceptual quality of 3D meshes. Additionally, to address the lack of research into content dependent image quality assessment, we propose an intelligent method of error scaling that depends on the scene content of the image. Finally we derive a low-level feature based method for improving the performance scores of LDR FR-IQA's by utilizing color features. In all cases, we compare our results with existing algorithms on multiple datasets and showed a clear improvement in performance.

We also present the results of a study that we performed on the analysis of surgical performance in laproscopic surgery emphasizing the differences in gaze patterns between expert and novice surgeons. A similar study can be used to explore the effects of human expertise with images on perceptual quality.

The studies that we presented represent the first exploration to incorporate dynamic range into blind IQA. The results could be expanded with additional studies into the perceptual quantities that we derived. Larger datasets can be used to improve the performance of this measure. Our results on content dependency were meant to demonstrate a proof of concept and can also be expanded using more complicated methods. A more complex scene identification system and more descriptive feature and be used to do this. Our results on the inclusion of color into IQA can be extended using a more continuous method of integrating features into IQA. Finally, our observations on texture quality in a tex-mesh can further be extended to include other factors like tessellation, material types etc.

References

- [1] W. X. A, L. Z. B, X. M. A, and A. C. B. C, *1 gradient magnitude similarity deviation: A highly efficient perceptual image quality index.* 86, 87
- [2] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bühlhoff, “Do hdr displays support ldr content?: A psychophysical evaluation,” *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 38, 2007. 23
- [3] M. M. Alam, K. P. Vilankar, D. J. Field, and D. M. Chandler, “Local masking in natural images: A database and analysis,” *Journal of Vision*, vol. 14, no. 8, pp. 22–22, Jul. 2014. DOI: 10.1167/14.8.22. 58, 62, 64
- [4] M. M. Alam, P. Patil, M. T. Hagan, and D. M. Chandler, “A computational model for predicting local distortion visibility via convolutional neural network trained on natural scenes,” in *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, Sep. 2015, pp. 3967–3971, ISBN: 978-1-4799-8339-1. DOI: 10.1109/ICIP.2015.7351550. [Online]. Available: <http://ieeexplore.ieee.org/document/7351550/>. 58, 64, 65
- [5] C. Amati, N. J. Mitra, and T. Weyrich, “A study of image colourfulness,” in *Proceedings of the Workshop on Computational Aesthetics - CAe '14*, New York, New York, USA: ACM Press, 2014, pp. 23–31, ISBN: 9781450330190. DOI: 10.1145/2630099.2630801. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2630099.2630801>. 68
- [6] T. O. Aydin, A. Smolic, and M. Gross, “Automated Aesthetic Analysis of Photographic Images,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 21, no. 1, pp. 31–42, DOI: doi:10.1109/TVCG.2014.2325047. [Online]. Available: <http://dx.doi.org/10.1109/TVCG.2014.2325047%20citeulike-article-id:14288802>. 21
- [7] T. O. Aydin, R. Mantiuk, and H.-P. Seidel, “Extending quality metrics to full luminance range images,” in *Electronic Imaging 2008*, International Society for Optics and Photonics, 2008, 68060B–68060B. 3, 21, 29, 34, 50, 51
- [8] H. Barlow and P. Földiák, “The computing neuron,” in, R. Durbin, C. Miall, and G. Mitchison, Eds., Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989, ch. Adaptation and Decorrelation in the Cortex, pp. 54–72, ISBN: 0-201-18348-X. [Online]. Available: <http://dl.acm.org/citation.cfm?id=103938.103942>. 86

- [9] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, Jan. 2013, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.89. [Online]. Available: <http://ieeexplore.ieee.org/document/6180177/>. 19
- [10] J.-Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm," Technical report, Intel Corporation Microprocessor Research Labs, 2000. 95
- [11] P. Breedveld and M. Wentink, "Eye-hand coordination in laparoscopy - an overview of experiments and supporting aids," *Minim Invasive Ther Allied Technol*, vol. 10, no. 3, pp. 155–162, 2001. 92
- [12] J. R. Cavanaugh, W. Bair, and J. A. Movshon, "Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons," *Journal of Neurophysiology*, vol. 88, no. 5, pp. 2547–2556, 2002. 81, 90
- [13] D. M. Chandler, M. D. Gaubatz, and S. S. Hemami, "A Patch-Based Structural Masking Model with an Application to Compression," *EURASIP Journal on Image and Video Processing*, vol. 2009, no. 1, pp. 1–22, 2009, ISSN: 1687-5176. DOI: 10.1155/2009/649316. [Online]. Available: <http://jivp.urasipjournals.com/content/2009/1/649316>. 65
- [14] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *Image Processing, IEEE Transactions on*, vol. 16, no. 9, pp. 2284–2298, 2007. 86, 87
- [15] H.-w. Chang, Q.-w. Zhang, Q.-g. Wu, and Y. Gan, "Perceptual image quality assessment by independent feature detector," *Neurocomputing*, vol. 151, pp. 1142–1152, 2015. 16, 86, 87
- [16] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep Image Saliency Computing via Progressive Representation Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016, ISSN: 2162-237X. DOI: 10.1109/TNNLS.2015.2506664. [Online]. Available: <http://ieeexplore.ieee.org/document/7372470/>. 20
- [17] I. Cheng and P. Boulanger, "A 3d perceptual metric using just-noticeable-difference," in *In Eurographics Short Presentations*, 2005, pp. 97–100. 24
- [18] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: Measuring error on simplified surfaces," Paris, France, France, Tech. Rep., 1996. 23

- [19] A. Coates, A. Y. Ng, and H. Lee, “An analysis of single-layer networks in unsupervised feature learning,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, 2011, pp. 215–223. [Online]. Available: <http://www.jmlr.org/proceedings/papers/v15/coates11a/coates11a.pdf>. 79
- [20] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, *Watermarked 3d mesh quality assessment*. 23
- [21] S. J. Daly, “Visible differences predictor: An algorithm for the assessment of image fidelity,” in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, International Society for Optics and Photonics, 1992, pp. 2–15. 16, 32
- [22] P. R. Delucia, “Griswold ja,” *Effects of camera arrangement on perceptual-motor performance in minimally invasive surgery .J Exp Psychol Appl.*, vol. 2011, no. 17, p. 3, DOI: 10.1037/a0024041. 93
- [23] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, “Adaptive logarithmic mapping for displaying high contrast scenes,” in *Computer Graphics Forum*, Wiley Online Library, vol. 22, 2003, pp. 419–426. 41
- [24] R. Dumont, F. Pellacini, and J. A. Ferwerda, “Perceptually-driven decision theory for interactive realistic rendering,” *ACM Trans. Graph.*, vol. 22, no. 2, pp. 152–181, Apr. 2003, ISSN: 0730-0301. DOI: 10.1145/636886.636888. [Online]. Available: <http://doi.acm.org/10.1145/636886.636888>. 23
- [25] D. Ellemborg, H. A. Allen, and R. F. Hess, “Second-order spatial frequency and orientation channels in human vision,” *Vision Research*, vol. 46, no. 17, pp. 2798–2803, 2006, ISSN: 0042-6989. DOI: <http://dx.doi.org/10.1016/j.visres.2006.01.028>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0042698906000629>. 80
- [26] F. D. Emin Zerman Giuseppe Valenzise, “An extensive performance evaluation of full-reference hdr image quality metrics,” in *Quality of Multimedia Experience (QoMEX) (Under 2nd review), year=2017*. 23, 26, 40, 41
- [27] D. R. Fuhrmann, J. A. Baro, and J. R. Cox Jr., “Experimental evaluation of psychophysical distortion metrics for jpeg-encoded images,” vol. 1913, 1993, pp. 179–190. DOI: 10.1117/12.152692. [Online]. Available: <http://dx.doi.org/10.1117/12.152692>. 12
- [28] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, “Saliency from hierarchical adaptation through decorrelation and variance normalization,” *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012. 86

- [29] B. Girod, "Digital images and human vision," in, A. B. Watson, Ed., Cambridge, MA, USA: MIT Press, 1993, ch. What's Wrong with Mean-squared Error? Pp. 207–220, ISBN: 0-262-23171-9. [Online]. Available: <http://dl.acm.org/citation.cfm?id=197765.197784>. 13
- [30] J. P. Guilford, "Psychometric methods," 1954. 103
- [31] G. B. Hanna and C. A. SurgEndosc, "Influence of the optical axis-to-target view angle on endoscopic task performance," *Surg EndoscApr*;, vol. 13, no. 4, pp. 371–5, 1999. 92, 93, 101
- [32] Hantao Liu, U. Engelke, Junle Wang, P. Le Callet, and I. Heynderickx, "How Does Image Content Affect the Added Value of Visual Attention in Objective Image Quality Assessment?" *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 355–358, Apr. 2013, ISSN: 1070-9908. DOI: 10.1109/LSP.2013.2243725. [Online]. Available: <http://ieeexplore.ieee.org/document/6423792/>. 21
- [33] Hantao Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, Jul. 2011, ISSN: 1051-8215. DOI: 10.1109/TCSVT.2011.2133770. [Online]. Available: <http://ieeexplore.ieee.org/document/5740313/>. 20, 76
- [34] D. Henderickx, K. Maetens, T. Geerinck, and E. Soetens, "Modeling the Interactions of Bottom-Up and Top-Down Guidance in Visual Attention," in, Springer, Berlin, Heidelberg, 2009, pp. 197–211. DOI: 10.1007/978-3-642-00582-4_{\ }15. [Online]. Available: http://link.springer.com/10.1007/978-3-642-00582-4_15. 5, 67, 74
- [35] E. Hildreth, "Theory of edge detection," *Proceedings of Royal Society of London*, vol. 207, no. 187-217, p. 9, 1980. 8, 16, 80
- [36] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 6, pp. 1275–1286, 2015. 37, 41
- [37] L. C. H.R. Sheikh Z.Wang and A. Bovik, "Live image quality assessment database release 2," 1, 26, 62, 65, 87
- [38] J. W. Hubber, N. Taffinder, and R. C. Russell, "Darzi a (2003)," *He effects of Different viewing conditions on performance in simulated minimal access surgery*, vol. 46, pp. 999–1016, 92, 93
- [39] B. Iglewicz and D. C. Hoaglin, "How to detect and handle outliers, asqc basic references in quality control," *Milwaukee, WI: American Society for Quality Control*, 1993. 106

- [40] N. Imamoglu, W. Lin, and Y. Fang, “A saliency detection model using low-level features based on wavelet transform,” *Multimedia, IEEE Transactions on*, vol. 15, no. 1, pp. 96–105, Jan. 2013, ISSN: 1520-9210. DOI: 10.1109/TMM.2012.2225034. 82
- [41] L. Itti, C. Koch, E. Niebur, *et al.*, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. [Online]. Available: citeulike-article-id:14288749. 19
- [42] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, “What is the best multi-stage architecture for object recognition?” In *Computer Vision, 2009 IEEE 12th International Conference on*, Sep. 2009, pp. 2146–2153. DOI: 10.1109/ICCV.2009.5459469. 79, 81
- [43] 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. 18, 40
- [44] ———, “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. [Online]. Available: citeulike-article-id:14288767. 31, 59, 60
- [45] ———, “Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks,” in *Image Processing (ICIP), 2015 IEEE International Conference on*, IEEE, 2015, pp. 2791–2795. 16, 36, 44
- [46] Y. Kao, R. He, and K. Huang, “Deep Aesthetic Quality Assessment with Semantic Information,” Apr. 2016. [Online]. Available: <http://arxiv.org/abs/1604.04970>. 21
- [47] Y. Kao, K. Huang, and S. Maybank, “Hierarchical aesthetic quality assessment using deep convolutional neural networks,” *Signal Processing: Image Communication*, vol. 47, no. C, pp. 500–510, Sep. 2016, ISSN: 09235965. DOI: 10.1016/j.image.2016.05.004. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0923596516300613>. 21
- [48] S. A. Karunasekera and N. G. Kingsbury, “Distortion measure for blocking artifacts in images based on human visual sensitivity,” vol. 2094, 1993, pp. 474–486. DOI: 10.1117/12.157966. [Online]. Available: <http://dx.doi.org/10.1117/12.157966>. 14
- [49] J. Kim, D. Han, Y.-W. Tai, and J. Kim, “Salient Region Detection via High-Dimensional Color Transform and Local Spatial Support,” *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 9–23, Jan. 2016, ISSN: 1057-7149. DOI: 10.1109/TIP.2015.2495122. [Online]. Available: <http://ieeexplore.ieee.org/document/7307162/>. 72

- [50] S.-J. Kim, S.-K. Kim, and C.-H. Kim, “Discrete differential error metric for surface simplification,” in *Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*, ser. PG '02, Washington, DC, USA: IEEE Computer Society, 2002, pp. 276–, ISBN: 0-7695-1784-6. [Online]. Available: <http://dl.acm.org/citation.cfm?id=826030.826623>. 23
- [51] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *ArXiv preprint arXiv:1412.6980*, 2014. 40
- [52] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, “Subjective quality assessment database of hdr images compressed with jpeg xt,” in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, IEEE, 2015, pp. 1–6. 26, 40, 41
- [53] N. K. Kottayil, G. Valenzise, F. Dufaux, and I. Cheng, “Blind Quality Estimation by Disentangling Perceptual and Noisy Features in High Dynamic Range Images,” *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1512–1525, Mar. 2018, ISSN: 1057-7149. DOI: 10.1109/TIP.2017.2778570. [Online]. Available: <http://ieeexplore.ieee.org/document/8123879/>. 60, 61
- [54] D. Kundu and B. L. Evans, “Visual attention guided quality assessment of tone-mapped images using scene statistics,” in *Image Processing (ICIP), 2016 IEEE International Conference on*, IEEE, 2016, pp. 96–100. 22
- [55] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, “No-reference image quality assessment for high dynamic range images,” in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, 2016. 22
- [56] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, “Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions,” *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, Jan. 1998, ISSN: 1052-6234. DOI: 10.1137/S1052623496303470. [Online]. Available: <http://epubs.siam.org/doi/10.1137/S1052623496303470>. 75
- [57] E. C. Larson and D. M. Chandler, “Most apparent distortion: A dual strategy for full-reference image quality assessment,” in *Proc. SPIE*, vol. 7242, 2009, 7, 16, 80, 86, 87
- [58] E. C. Larson and D. Chandler, “Categorical image quality (csiq) database,” *Online*, <http://vision.okstate.edu/csiq>, 2010. 1, 26, 87
- [59] E. Larson and D. Chandler, “Most apparent distortion: a dual strategy for full-reference image quality assessment,” in *Proc. SPIE*, vol. 7242, 2009. [Online]. Available: [citeulike-article-id:1428856](http://dx.doi.org/10.1117/12.88856). 19, 62, 65, 69, 75

- [60] G. Lavoué, “A local roughness measure for 3d meshes and its application to visual masking,” *ACM Trans. Appl. Percept.*, vol. 5, no. 4, 21:1–21:23, Feb. 2009, ISSN: 1544-3558. DOI: 10.1145/1462048.1462052. [Online]. Available: <http://doi.acm.org/10.1145/1462048.1462052>. 23
- [61] —, “A multiscale metric for 3d mesh visual quality assessment,” in *Computer Graphics Forum*, Wiley Online Library, vol. 30, 2011, pp. 1427–1437. 24
- [62] G. Lavoué, I. Cheng, and A. Basu, “Perceptual quality metrics for 3d meshes: Towards an optimal multi-attribute computational model,” in *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, ser. SMC '13, Washington, DC, USA: IEEE Computer Society, 2013, pp. 3271–3276, ISBN: 978-1-4799-0652-9. DOI: 10.1109/SMC.2013.557. [Online]. Available: <http://dx.doi.org/10.1109/SMC.2013.557>. 24
- [63] G. Lavoue, E. D. Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, “Perceptually driven 3d distance metrics with application to watermarking,” in *SPIE Optics+ Photonics*, International Society for Optics and Photonics, 2006, pp. 63120L–63120L. 24
- [64] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 2169–2178. [Online]. Available: [citeulike-article-id:14288745](http://dx.doi.org/10.1109/CVPR.2006.14288745). 71
- [65] P. Le Callet and F. Autrusseau, *Subjective quality assessment irccyn/ivc database*, Subjective database, 2005. 26
- [66] G. Li and Y. Yu, “Visual Saliency Detection Based on Multiscale Deep CNN Features,” *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016, ISSN: 1057-7149. DOI: 10.1109/TIP.2016.2602079. [Online]. Available: <http://ieeexplore.ieee.org/document/7548372/>. 19
- [67] A. Liu, W. Lin, and M. Narwaria, “Image quality assessment based on gradient similarity,” *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012, ISSN: 1057-7149. DOI: 10.1109/TIP.2011.2175935. 16, 86, 87
- [68] H. Liu and I. Heynderickx, “Towards an efficient model of visual saliency for objective image quality assessment,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, Mar. 2012, pp. 1153–1156. DOI: 10.1109/ICASSP.2012.6288091. 82
- [69] L. Liu, B. Liu, H. Huang, and A. C. Bovik, “No-reference image quality assessment based on spatial and spectral entropies,” *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014. 17, 18, 40, 41

- [70] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: [citeulike-article-id:14288819](#). 71
- [71] J. L. Mannos and D. J. Sakrison, “The Effects of a Visual Fidelity Criterion on the Encoding of Images,” *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, Jul. 1974. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs%5C_all.jsp?arnumber=1055250. 12
- [72] R. Mantiuk, S. Daly, and L. Kerofsky, “Display adaptive tone mapping,” in *ACM Transactions on Graphics (TOG)*, ACM, vol. 27, 2008, p. 68. 41
- [73] 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. 17, 39, 40, 44, 45
- [74] M. Miyahara, K. Kotani, and V. Algazi, “Objective picture quality scale (pqs) for image coding,” *Communications, IEEE Transactions on*, vol. 46, no. 9, pp. 1215–1226, Sep. 1998, ISSN: 0090-6778. DOI: 10.1109/26.718563. 14
- [75] A. K. Moorthy and A. C. Bovik, “A two-step framework for constructing blind image quality indices,” *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 513–516, 2010. 17, 40
- [76] ———, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3350–3364, 2011. 17, 40
- [77] A. Moorthy and A. Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3350–3364, 2011. [Online]. Available: [citeulike-article-id:14288763](#). 61
- [78] H. Z. Nafchi, A. Shahkolaei, R. F. Moghaddam, and M. Cheriet, “Fsimt: A feature similarity index for tone-mapped images,” *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1026–1029, 2015. 22
- [79] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pépion, “Tone mapping-based high-dynamic-range image compression: Study of optimization criterion and perceptual quality,” *Optical Engineering*, vol. 52, no. 10, pp. 102 008–102 008, 2013. 25, 40, 41
- [80] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pépion, “Impact of tone mapping in high dynamic range image compression,” in *VPQM*, 2014, pp–1. 26, 40, 41
- [81] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, “Hrdvdp-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images,” 1, vol. 24, International Society for Optics and Photonics, 2015, pp. 010 501–010 501. 22, 29, 32, 34

- [82] N. B. Nill and B. Bouzas, "Objective image quality measure derived from digital image power spectra," *Optical Engineering*, no. 4, pp. 813–825, 1992. DOI: 10.1117/12.56114. [Online]. Available: <http://dx.doi.org/10.1117/12.56114>. 12
- [83] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001, ISSN: 09205691. DOI: 10.1023/A:1011139631724. [Online]. Available: <http://link.springer.com/10.1023/A:1011139631724>. 71
- [84] O. Olmos, C. D. Wickens, and A. Chudy, "Tactical displays for combat awareness: An examination of dimensionality and frame of reference concepts and the application of cognitive engineering," *The International Journal of Aviation Psychology*, vol. 10, pp. 247–371, 2000. 93
- [85] T. P. O’rourke and R. L. Stevenson, "Human visual system based wavelet decomposition for image compression," *J. VCIP V*, vol. 6, pp. 109–121, 1995. 13
- [86] X. Otazu, M. Vanrell, and C. A. Párraga, "Multiresolution wavelet framework models brightness induction effects," *Vision Research*, vol. 48, no. 5, pp. 733–751, 2008, ISSN: 0042-6989. DOI: <http://dx.doi.org/10.1016/j.visres.2007.12.008>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0042698907005597>. 81
- [87] Y. Pan, I. Cheng, and A. Basu, "Quality metric for approximating subjective evaluation of 3-d objects," *Trans. Multi.*, vol. 7, no. 2, pp. 269–279, Apr. 2005, ISSN: 1520-9210. DOI: 10.1109/TMM.2005.843364. [Online]. Available: <http://dx.doi.org/10.1109/TMM.2005.843364>. 6, 24, 103, 106, 107, 109
- [88] E. Peli, "Contrast in complex images," *J. Opt. Soc. Am. A*, vol. 7, no. 10, pp. 2032–2040, Oct. 1990. DOI: 10.1364/JOSAA.7.002032. [Online]. Available: <http://josaa.osa.org/abstract.cfm?URI=josaa-7-10-2032>. 13
- [89] 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. 26, 87
- [90] —, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015. [Online]. Available: [citeulike-article-id:14288787](http://dx.doi.org/10.1016/j.spi.2015.04.008). 62, 65
- [91] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009. 26, 87

- [92] M. Ramasubramanian, S. N. Pattanaik, and D. P. Greenberg, “A perceptually based physical error metric for realistic image synthesis,” in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., 1999, pp. 73–82. 3
- [93] E. Reinhard and K. Devlin, “Dynamic range reduction inspired by photoreceptor physiology,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, pp. 13–24, Jan. 2005, ISSN: 1077-2626. DOI: 10.1109/TVCG.2005.9. 41
- [94] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographic tone reproduction for digital images,” *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3, pp. 267–276, 2002. 41
- [95] A. M. Rohaly, P. J. Corriveau, J. M. Libert, A. A. Webster, V. Baroncini, J. Beerends, J.-L. Blin, L. Contin, T. Hamada, D. Harrison, A. P. Hekstra, J. Lubin, Y. Nishida, R. Nishihara, J. C. Pearson, A. F. Pessoa, N. Pickford, A. Schertz, M. Visca, A. B. Watson, and S. Winkler, *Video quality experts group: Current results and future directions*, 2000. DOI: 10.1117/12.386632. [Online]. Available: <http://dx.doi.org/10.1117/12.386632>. 86
- [96] N. Sadaka, L. Karam, R. Ferzli, and G. Abousleman, “A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling,” in *2008 IEEE International Conference on Image Processing, ICIP 2008*, 2008. [Online]. Available: [citeulike-article-id:14288746](http://dx.doi.org/10.1109/ICIP.2008.4466346). 21
- [97] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, “Complex wavelet structural similarity: A new image quality index,” in *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 2009. 15, 86, 87
- [98] H. Sheikh, A. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” *Image Processing, IEEE Transactions on*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005, ISSN: 1057-7149. DOI: 10.1109/TIP.2005.859389. 86, 87
- [99] Sheikh and A. C. Bovik, “LIVE Image Quality Assessment Database Release 2,” [Online]. Available: [citeulike-article-id:14288786](http://live.ece.utd.edu/). 20
- [100] J. Shi and C. Tomasi, “Good features to track,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR’94)*, , Seattle, Jun. 1994, pp. 593–600. 96
- [101] E. Siahaan, A. Hanjalic, and J. A. Redi, “Does visual quality depend on semantics? A study on the relationship between impairment annoyance and image semantics at early attentive stages,” *Electronic Imaging*, vol. 2016, no. 16, pp. 1–9, Feb. 2016, ISSN: 2470-1173. DOI: 10.2352/ISSN.2470-1173.2016.16.HVEI-109. [Online]. Available: <http://dx.doi.org/10.2352/ISSN.2470-1173.2016.16.HVEI-109>.

//www.ingentaconnect.com/content/10.2352/ISSN.2470-1173.2016.16.HVEI-109.

5, 68, 69

- [102] P. Teo and D. Heeger, "Perceptual image distortion," in *Proceedings of 1st International Conference on Image Processing*, vol. 2, IEEE Comput. Soc. Press, pp. 982–986, ISBN: 0-8186-6952-7. DOI: 10.1109/ICIP.1994.413502. [Online]. Available: <http://ieeexplore.ieee.org/document/413502/>.
- [103] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.
- [104] Z. Ugray, L. Lasdon, J. Plummer, F. Glover, J. Kelly, and R. Martí, "Scatter Search and Local NLP Solvers: A Multistart Framework for Global Optimization," *INFORMS Journal on Computing*, vol. 19, no. 3, pp. 328–340, Aug. 2007, ISSN: 1091-9856. DOI: 10.1287/ijoc.1060.0175. [Online]. Available: <http://pubsonline.informs.org/doi/10.1287/ijoc.1060.0175>.
- [105] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for hdr image compression," in *SPIE Optical Engineering+ Applications*, International Society for Optics and Photonics, 2014, pp. 92170C–92170C.
- [106] L. Váša and J. Rus, "Dihedral angle mesh error: A fast perception correlated distortion measure for fixed connectivity triangle meshes," *Computer Graphics Forum*, vol. 31, no. 5, pp. 1715–1724, 2012, ISSN: 1467-8659. DOI: 10.1111/j.1467-8659.2012.03176.x. [Online]. Available: <http://dx.doi.org/10.1111/j.1467-8659.2012.03176.x>.
- [107] S. J. Vine, R. S. W. Masters, J. S. McGrath, E. Bright, and M. R. Wilson, "Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills, surgery, volume 152, issue 1, july 2012," vol. 32, pp. 39–6060,
- [108] 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.
- [109] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. [Online]. Available: [citeulike-article-id:14288857](http://dx.doi.org/10.1109/TIP.2004.14288857).

65

38

75

26, 40, 41

23

93

7, 12, 15, 16, 24, 80, 82,

58, 75

- [110] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," *Trans. Img. Proc.*, vol. 20, no. 5, pp. 1185–1198, DOI: doi : 10 . 1109 / TIP . 2010 . 2092435. [Online]. Available: <http://dx.doi.org/10.1109/TIP.2010.2092435>%20citeulike-article-id:14288830. 19
- [111] 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*, Ieee, vol. 2, 2003, pp. 1398–1402. 15, 22, 86, 87
- [112] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking.," *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 14, no. 9, pp. 2379–91, Sep. 1997, ISSN: 1084-7529. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9291608>. 65
- [113] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images". 60
- [114] S. Westen, R. Lagendijk, and J. Biemond, "Perceptual image quality based on a multiple channel hvs model," in *Proceedings of ICASSP*, 1995, pp. 2351–2354. 13
- [115] C. D. Wickens, L. C. Thomas, and R. Young, "Frames of reference for the display of battlefield information: Judgment-display dependencies," *Human Factors*, vol. 42, pp. 660–675, 2000. 93
- [116] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *Image Processing, IEEE Transactions on*, vol. 22, no. 1, pp. 43–54, Jan. 2013, ISSN: 1057-7149. DOI: 10.1109/TIP.2012.2214048. 16, 87
- [117] J. Wu, F. Qi, and G. Shi, "Self-similarity based structural regularity for just noticeable difference estimation," *Journal of Visual Communication and Image Representation*, vol. 23, no. 6, pp. 845–852, 2012, ISSN: 1047-3203. DOI: <http://dx.doi.org/10.1016/j.jvcir.2012.04.010>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1047320312000788>. 14
- [118] W. Xue, L. Zhang, X. Mou, and Alan, *1 Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index*. [Online]. Available: citeulike-article-id:14288784. 75
- [119] J. Yang and M.-H. Yang, "Top-Down Visual Saliency via Joint CRF and Dictionary Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 576–588, Mar. 2017, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2547384. [Online]. Available: <http://ieeexplore.ieee.org/document/7442536/>. 19

- [120] A. L. Yarbus, “Eye Movements During Perception of Complex Objects,” in *Eye Movements and Vision*, Boston, MA: Springer US, 1967, pp. 171–211. DOI: 10.1007/978-1-4899-5379-7_{_}8. [Online]. Available: http://link.springer.com/10.1007/978-1-4899-5379-7_8. 20
- [121] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Real-time no-reference image quality assessment based on filter learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 987–994. 18
- [122] H. Yeganeh and Z. Wang, “Objective quality assessment of tone-mapped images,” *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, 2013. 22
- [123] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux, “The relation between MOS and pairwise comparisons and the importance of cross-content comparisons,” in *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2018)*, 2018. 62
- [124] L. Zhang, Y. Shen, and H. Li, “VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment,” *Image Processing, IEEE Transactions on*, vol. 23, no. 10, pp. 4270–4281, DOI: doi:10.1109/TIP.2014.2346028. [Online]. Available: <http://dx.doi.org/10.1109/TIP.2014.2346028%20citeulike-article-id:14288806>. 19
- [125] —, “Vsi: A visual saliency-induced index for perceptual image quality assessment,” *Image Processing, IEEE Transactions on*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014, ISSN: 1057-7149. DOI: 10.1109/TIP.2014.2346028. 16, 86, 87
- [126] L. Zhang, D. Zhang, X. Mou, and D. Zhang, “FSIM: A Feature Similarity Index for Image Quality Assessment,” *Image Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2378–2386, DOI: doi:10.1109/TIP.2011.2109730. [Online]. Available: <http://dx.doi.org/10.1109/TIP.2011.2109730%20citeulike-article-id:14288822>. 19, 75
- [127] L. Zhang, D. Zhang, X. Mou, and D. Zhang, “Fsim: A feature similarity index for image quality assessment,” *Image Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011, ISSN: 1057-7149. DOI: 10.1109/TIP.2011.2109730. 16, 86, 87, 89
- [128] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, “The Application of Visual Saliency Models in Objective Image Quality Assessment: A Statistical Evaluation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016, ISSN: 2162-237X. DOI: 10.1109/TNNLS.2015.2461603. [Online]. Available: <http://ieeexplore.ieee.org/document/7185444/>. 21

- [129] T. Zhu and L. Karam, “A no-reference objective image quality metric based on perceptually weighted local noise,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–8, 2014.

18, 32, 34, 58, 60