Subjective and Objective Evaluation of Tone-Mapping and De-Ghosting Algorithms

Manchana Akshai Krishna

A Thesis Submitted to
Indian Institute of Technology Hyderabad
In Partial Fulfillment of the Requirements for
The Degree of Master of Technology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Electrical Engineering

July 2016

Declaration

I declare that this written submission represents my ideas in my own words, and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

M. Akshai Krishna

(Signature)

Manchana Akshai Krishna

(– Student Name –)

EE13M1029

(Roll No)

Approval Sheet

This Thesis entitled Subjective and Objective Evaluation of Tone-Mapping and De-Ghosting Algorithms by Manchana Akshai Krishna is approved for the degree of Master of Technology from IIT Hyderabad

P. Rajal

Dr. P.Rajalakshmi
Associate Professor, Department of Electrical Engineering, IIT Hyderabad

Dr. Ketan Detroja
Associate Professor, Department of Electrical Engineering, IIT Hyderabad

Sathya /08.07.2016

Dr. Sathya Peri
Associate Professor, Department of Computer Science Engineering, IIT Hyderabad

8/7/16

Dr. Sumohana Channappayya
Assistant Professor, Department of Electrical Engineering, IIT Hyderabad
Adviser

Abstract

With the increasing importance of high dynamic range (HDR) imaging and low availability of HDR displays, HDR cameras the need for efficient tone mapping, De-ghosting techniques is very crucial. However the tone mapping operators, De-Ghosting tend to introduce distortions in the HDR images, thus making it visually unpleasant in normal displays. Subjective evaluation of images is important for rating these algorithms as the users should be able to visualize the complete details present in both the brightly and poorly illuminated regions of the scene. To facilitate a systematic subjective study we have created a database of HDR images tone mapped, De-Ghosted using popular algorithms. We conducted a subjective study of the tone mapped images, computed objective scores by using some of the state-of-the-art no-reference low dynamic range image quality assessment algorithms and evaluated their performance. We show that a moderate and low correlation between objective and subjective scores indicates the need for the consideration of human perception in rating tone mapping operators and De-Ghosting algorithms.

Contents

Chapter 1

Introduction

High dynamic range (HDR) images have diverse applications ranging from remote sensing, global illumination, image based modeling, to virtual reality. HDR image can be captured by using specially manufactured HDR camera or by fusion of multiple HDR images. HDR imaging is gaining popularity even in high end mobile devices because of the ease in the generation using multi-exposure images and the amount of scene details depicted with higher contrast. These images have very high contrast in both brightly and poorly illuminated regions of the scene. The special displays for viewing the HDR images are very expensive and consume a lot of power and space. HDR images cannot be displayed on popularly used LDR displays. Hence, we require algorithms to convert HDR images to LDR images for display purposes. This process of converting HDR images to LDR images is called tone mapping or tone reproduction. The operator which performs this task is known as a tone mapping operator (TMO). For reproduction of natural scenes, visual models inspired by human visual system (HVS) are used for tone mapping. The state-of-the-art tone mapping techniques are reviewed in the books - [1] and [2]. The HDR image has a significance if we can render these images on LDR displays. The LDR representation of the HDR image which can be used by mostly used devices is the final step of any algorithm. So the HDR image generation itself has no importance. The algorithms might directly go to the LDR generation step without involving the generation of HDR. So while formulating algorithms to formulate a HDR image by using multiple LDR image fusion we can eliminate the HDR generation step directly. The generation of HDR image by using multiple LDR images assumes the scene to be static. The real world applications of HDR image include alot dynamic scenes with a lot of scene changes and camera motion. This causes unnatural artifacts called as GHOSTs. These are the replicated content present in the output of the algorithms. These GHOSTs causes reduction of scene details and affect the perception of the humans. The algorithms which improve the visual details by using multiple LDR image fusion and minimize the GHOSTs generated are called as De-Ghosting algorithms. The evaluation of these De-Ghosted images is required to find the performance of De-Ghosting algorithms and report the best working parameters for the construction of De-Ghosting algorithms and Objective algorithms to evaluate the performance of these De-Ghosting algorithms.

The contrast in real world scenes and in the captured images using common digital cameras differ a lot. Typical natural scenes have very high contrast based on their illumination but due to limited sensor well capacity and limited bit depth of images, we cannot capture the entire dynamic range of the natural scene. The pixel values of HDR images are modified by a tone mapping operator in such a way that the local contrast variations look realistic and rich with details. However, this process may introduce different types of distortions. Different tone mapping operators affect the HDR images differently such as altering the maximum luminance value, gradients, edges, etc.

There is a need for evaluating these tone mapped images on the basis of distortions present in them which affect the human perception. A majority of the current state-of-the-art objective quality assessment algorithms consider gray scale images that do not include colour tone mapped images. This creates a need for perceptually better method for measuring colour information. The current state-of-the-art full reference dynamic range independent metric uses contrast as the main metric to find objective quality score [3]. However, this metric does not consider the effect of contrast modification on human perception for given scene statistics. This calls for a perceptually motivated quality assessment algorithm. This motivates the use of a subjective quality assessment which can direct the tone mapping operators to be adapted based on contents of the scene. The subjective ratings specify the human response to contrast modifications for a particular scene. This provides a ground truth for objective evaluation scores to match the corresponding subjective scores. The tone mapping operator optimisaton algorithm [4] showed the significance of considering human perception in tone mapping and validated their results by using subjective evaluation. This motivates the need for subjective evaluation and its consideration in evaluating the performance of tone mapping operators.

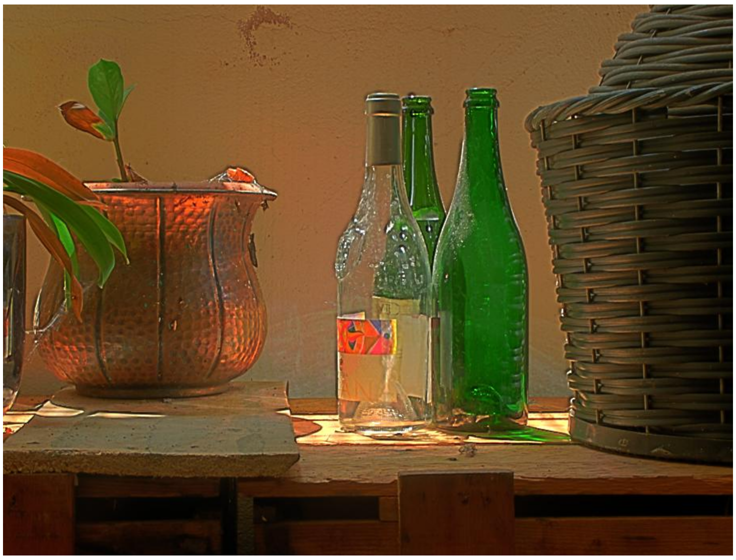


Figure 1.1: Ashikhmin TMO[5]

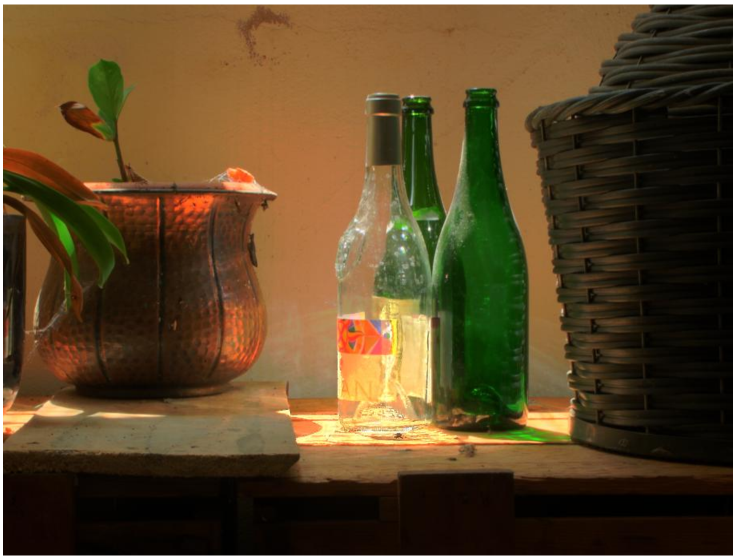


Figure 1.2: Banterle TMO [6]

Chapter 2

Subjective Evaluation of Tone-Mapping Algorithms

We selected a set of 44 HDR images of natural scenes a wide dynamic range gathered from various sources in Radiance RGBE (.hdr) format. In this subjective study, instead of using 44 images which would be a strenuous task, we selected 20 with a good spread of contrast values. The database of tone mapped images is constructed using 21 state-of-the-art tone mapping operators that are reviewed in [1]. This leads to the generation of 420 LDR images. A total of 21 users consisting of 19 male and 2 female between 20-40 years of age gave ratings. We did not show the reference HDR image to the subjects while rating was being done.

2.1 Soting and Database

The contrast of an image is the ratio of maximum luminance to the minimum luminance values represented in the CIE-Lab space. Dynamic range D of an image can be computed as:

$$D = 20 \log_{10} \left(\frac{L_{\max}}{(L_{\min} + c)} \right). \tag{2.1}$$

The constant added in the denominator c is 0 for nonzero values of $L_{\min}$ and 10^{-6} for zero minimum luminance images. Experimentally we observed that, there are significant number of pixels in all the HDR images with luminance value less than 10^{-6} in the zero minimum luminance condition. We found the dynamic range for the entire set of 44 HDR images with values ranging from 39.9110dB to 159.99dB. Based on the values of dynamic range we sorted these images into three categories namely high dynamic range images, medium dynamic range images, low dynamic range images. Then we randomly selected equal number of images from each category and formed a set of 20 images which had a wide spread of contrast values. These set of images thus had different illumination conditions, varying content and textures. We used 21 tone mapping operators, [5], [6] [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] [22], with default input parameters. This choice of parameters has been reported in literature to produce the best tone mapping result for a particular operator. Thus, a total of 420 tone mapped images were created from 20 HDR images.

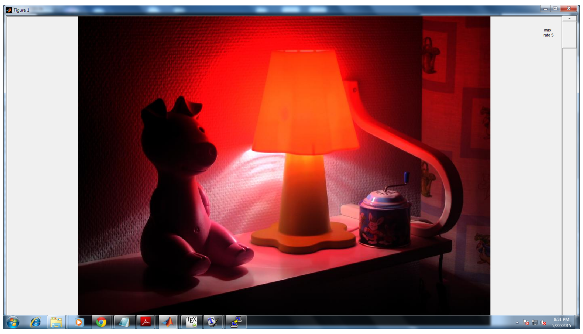


Figure 2.1: Graphical User Interface.

2.2 Experimental Setup

The evaluation was conducted in a laboratory on a computer with 55.88 cm LCD ACER monitor running Windows 7 operating system. The screen was set to resolution of 1366 × 768.The images have maximum and minimum vertical resolution of 2272, 525 and horizontal resolution of 2272, 768. We used MATLAB 2013b for creating graphical user interface. The subjects viewed the images on the monitor from a distance of 75 cm. The brightness of the display was set to be the maximum possible brightness in Window 7 system.

Every subject rated 420 images. For convenience, we created a user interface as specified in [23], which was simple and easy to use. We used a continuous scale from 0-5, 0 for minimum rating and 5 for maximum rating. We used a vertical slider at the right end of the monitor ranging from 0 to 5 as shown in Fig.2.1. With one mouse click we can slide the slider to the required rating position. To go to the next image, the subject had to touch/click on the image. The maximum rating and corresponding slider position were mentioned at the top right corner of the GUI. For giving maximum rating and going to the next image the subject had to double tap on the image window. For each user, it took an average of 30 minutes to evaluate which is an acceptable period before pausing the subjective evaluation process according to [23]. So, each subject completed evaluation in a single time window. The images were randomly permuted for each user before showing the set of images to the subjects. To avoid hysteresis we randomized the sequence of images for each subject.

2.3 Removal of Outliers

The process of evaluating 420 images is a very difficult and time consuming task. The subjects have the possibility of losing concentration or sliding the rating incorrectly. These ratings should not be considered while finding the mean subjective scores for each image. The subjects with large number of incorrect ratings should be discarded. We followed regulations specified in [23] for eliminating bad subjects and discarding any outliers.

According to the procedure followed in [23], the subject can be accepted if he/she has 95 percent or more in acceptable range. The acceptable range depends upon the distribution of the ratings. First we determined if the ratings follow normal distribution by observing the variance and kurtosis of the subjective ratings. Then the equations below specified in [23] gives the acceptable range $mean_i \geq \overline{u_i} \pm C \times Var$for accepting the rating of an user for an image. The parameter C will be selected by finding the distribution of the subjective scores.

$$mov_x = \frac{\sum_{i=1}^{N}(u_i - \overline{u_i})}{N}, \beta_2 = \frac{mov_4}{(mov_2)^2}$$

When we applied the above algorithm 111 ratings out of 8820 ratings given by 21 different subjects were marked as incorrect ratings. 3 users had zero errors. The maximum number of errors were 20. According to the above specified recommendation the maximum acceptable errors are 21 which is 5% of total number of 420 images. So all the 21 subject ratings were considered to be valid. While calculating the mean behaviour of the subjects for all the 420 image set, these 111 incorrect observations were not considered.

Chapter 3

Objective evaluation of Tone-Mapping Algorithms

3.1 Selection of Objective Metric

All the subjective scores were obtained in the range 0-5. 5 corresponds to best image with low perceivable distortions and 0 corresponds to the worst image with a lot of visible distortions. The slider will give the values with an accuracy of 6 decimal places. We tabulated three of the most and least correlating TMOs in Table.3.1. The Fig.3.1, Fig.3.2, Fig.3.3, Fig.3.5 show the scatter plots of subjective scores with NIQE, no-reference, and full reference objective metric scores respectively. The subjective scores given for ReinhardTMO were the best mean score of 0.5247. WardGlobalTMO was given the lowest subjective rating 1.6436. Since this study involved only LDR images, we computed only the Mean Opinion Score (not DMOS). The mean subjective scores which were in range 1.6 to 3.6 suggesting that the tone mapping operators are failing for some of the images and not producing perceptually best images. As we included a large variety of illuminations and natural scenes, the mean scores given were in the order of 3. This provides an opportunity to verify which tone mapping operators are working well for the type of illumination and scene.

3.2 Objective Scores

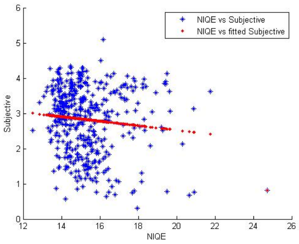


Figure 3.1: Scatter plot of subjective scores and NIQE.

We used state-of-the-art no-reference objective quality metrics QAC [24], SBIQE [25], NIQE [26],

TMO	Subjective	PSNR	FSIM	QAC	SBIQE	NIQE	PSNRCC	FSIMCC	QACCC	SBIQCC	NIQECC
KimKautzConsistentTMO	3.6832	5.3548	2.8465	3.4998	3.2127	4.2585	0.2927	0.2496	0.1323	0.3399	0.2962
ReinhardTMO	3.5247	6.6529	2.9493	3.4476	3.0984	4.2573	0.33332	0.2857	0.0331	0.1714	0.3729
LischinskiTMO	3.4748	6.5881	2.7859	3.4741	3.3314	4.2614	0.2717	0.1714	0.2286	0.2135	0.3609
WardGlobalTMO	1.6436	10.5537	4.2058	2.5781	2.1389	4.2102	0.3315	0.5172	0.5398	0.6602	0.5984
TumblinRushmeierTMO	1.7776	8.8751	4.0185	3.1068	2.4767	4.24823	0.3884	0.5308	0.1729	0.5143	0.4135
NormalizeTMO	1.8443	9.3288	3.9996	2.5688	2.0817	4.2065	0.3597	0.6647	0.6226	0.5444	0.4677

Table 3.1: Correlation with Objective scores.

Algorithm	Subject1	Subject2	Subject3	Subject4	Subject5	Subject6	Subject7	Subject8	Subject9	mean
QAC LCC	0.2024	0.3344	0.3184	0.1862	0.3517	0.4129	0.3770	0.5761	0.2652	0.6008
SBIQA LCC	0.2421	0.3764	0.3773	0.1386	0.3271	0.4057	0.3595	0.5503	0.2592	0.5336
NIQE LCC	0.0094	0.1586	0.1577	0.0763	0.1532	0.2069	0.1395	0.1794	0.0229	0.2054
FSIM LCC	0.3158	0.4753	0.4256	0.1803	0.4109	0.5325	0.4032	0.6544	0.2645	0.6973

Table 3.2: Subjective scores correlation with QAC, SBIQE, NIQE, and FSIM.

full reference objective quality metrics FSIM [27] and PSNR for objective evaluation. The objective scores were in the range of 2 to 4.3. NIQE produced quality scores between 4.18 and 4.27. The QAC produced scores between 2.5 and 3.6. The SBIQE produced scores between 2 and 3.8. NIQE scores showed very little variations around 4.2. QAC and SBIQE lead to scores similar to the range of subjective scores with equal amount of spread.

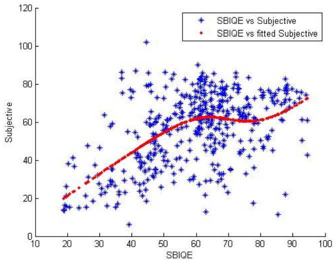


Figure 3.2: SBIQE

FSIM considers structural similarity between the images. In our experiment we have seen that the maximum mean objective score produced by FSIM was 3.9676 for FerwerdaTMO. But the subjective mean opinion score for FerwerdaTMO is 1.8786 which was one of the least mean opinion scores given by the subjects. This suggests that structural similarity was not that important to subjects within an acceptable range.The FSIM also compared well for other TMOs. This suggests that there is a need for learning the human behaviour to certain changes in structure of the tone mapped images, to rate that change good or bad where the subjective ratings are quite significant.

PSNR is a full reference metric which considers the mean squared error between reference and test image as a metric for objective evaluation. The maximum PSNR value which implies minimum mean squared error(MSE) was 12.0601 for LogarithmicTMO. But the mean subjective score for this TMO (2.0214) was one of the least subjective scores. ExponentialTMO(least PSNR 0.9111) had MOS score of 2.6751 which was not the least. The mean correlation of PSNR for all the TMOs on the entire database was 0.2915. We can infer that MSE does not play an important role in evaluating the TMOs.

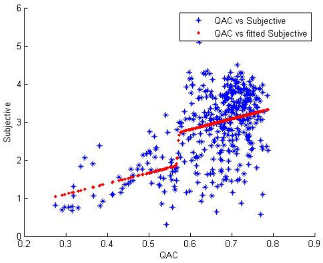
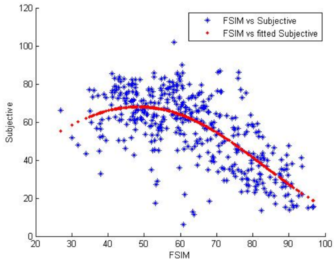


Figure 3.3: QAC



S

Figure 3.4: FSIM

3.3 Conclusions of Subjective and Objective Evaluation of Tone-Mapping Algorithms

3.3.1 Objective Evaluation

Table3.2 includes objective score correlation coefficients with nine subjects and the MOS values. Fig.3.6 shows the mean correlations of the 21 TMO subjective scores with their respective objective scores. The no-reference objective quality assessment metric QAC lead to highest correlation (0.6) with the subjective scores which is the best after the full reference FSIM. It was giving highest correlation of 0.8131 for the set AshikhminTMO and lowest correlated with BanterleTMO with correlation coefficient of 0.0125. AshikhminTMO was given ratings with mean subjective score of 3.4 which was one of the high subjective scores but not the best. But the main observation was the QAC was least correlated with the best KimKautzConsistentTMO which has highest mean opinion score of 3.7 having a correlation coefficient of 0.1. This suggests that the modifications done to HDR image by KimKautzConsistentTMO were marked to be bad by QAC. We can consider these changes which were accepted by the subjects as well.

SBIQE was also doing reasonably well with mean correlation of 0.5336. ReinhardBilTMO which was one of the better TMOs gave least correlation having correlation coefficient of 0.0913. SBIQE was highly correlated with WardGlobalTMO which was given lowest subjective subjective scores. Overall the SBIQE was working well for large number of TMOs. From these observations, we can pick some of the metrics followed by the objective algorithms. We might be able to suggest the best TMO if we can learn natural scene statistics of HDR images and classify them for the usage of TMO by using these subjective scores.

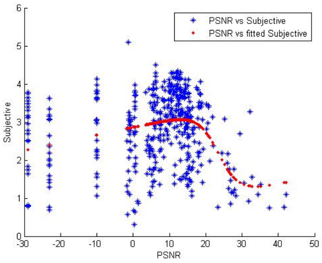


Figure 3.5: PSNR

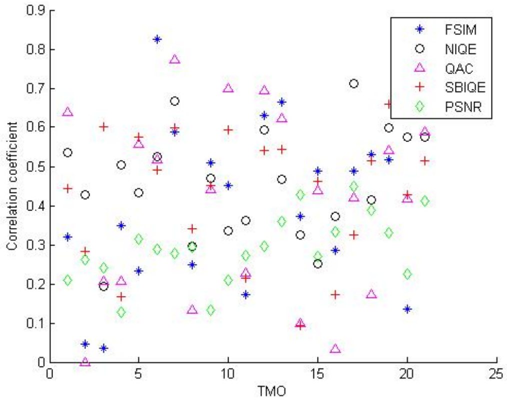


Figure 3.6: Scatter plots of correlation coefficients of subjective and objective scores.

3.3.2 Conclusions

There are state-of-the-art full reference objective quality assessment metrics to rate tone mapped LDR images like [3] and HDR-VDP [28]. But there is no subjective data available to validate the output results. This provides us with an opportunity to evaluate the perceptual performance of these state-of-the-art metrics. The state-of-the-art full reference metric involving learning is HDR-VDP2 [29] which provides subjective ratings for only HDR images not the tone mapped images. By using ratings for these LDR tone mapped images, we can learn the human reaction to changes in the contrast.

By observing the subjective scores, overall correlation coefficients and correlations with different TMOs, we can say that the parameters used in state-of-the-art objective algorithms can still be used for ranking the tone mapping operators, but we will have to modify them and pick the best possible metrics with respect to subjective scores in order to improve the performance. We will be updating this database in future and it will be freely available for research groups. The following link http://iith.ac.in/∼lfovias contains sample tone mapped images used in this study. The details regarding the study are published in [30].

Chapter 4

Formulation of Objective Quality Metric

The objective evaluation of the TMOs motivated us to formulate a new objective quality metric to assess the performance of the Tone-Mapping algorithms. The moderate correlations of the full reference and no reference and algorithms suggests that these algorithms are failing in some cases. The scattering of the full reference and no reference algorithm has a wide spread. This shows that full reference and no reference algorithms were finding different information while assessing these algorithms. So if we can capture more information by combining the parameters of these no reference and full reference algorithms we can form a good Objective Quality Assessment metric.

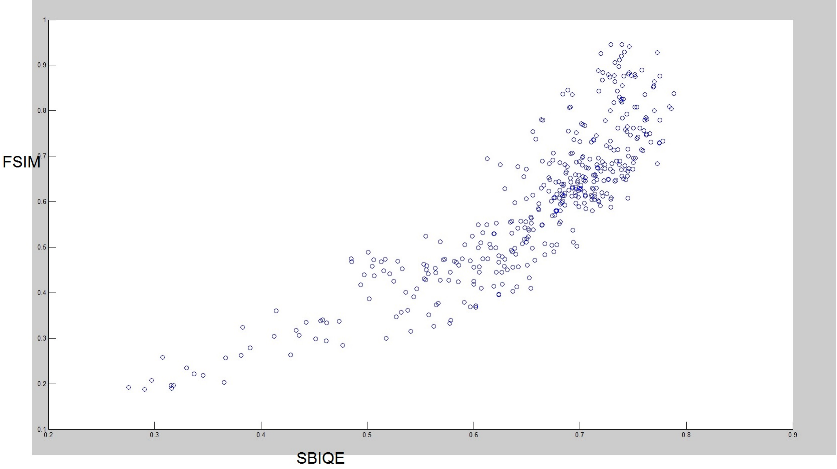


Figure 4.1: SBIQE vs FSIM

4.1 Challenges

The objective quality metric has various challenges like naturalness of the colour information present in the algorithm output, the structural modification of the reference image. There are very high amount of unnatural colour artifacts present in the algorithms outputs. Most of the current state of the art algorithms wont include colour information in assessing the quality. The subjective evaluation was no reference but, the reference image is present for each algorithm output. Although the subjects don't have information about the structure of the reference image, by their learning the subjects have a good knowledge of natural structures. With observation we can say that, while assessing these algorithms we can use the structural information because a good algorithm should preserve as much structural information as possible. So the objective quality metric should be full reference. But even the structure of the information is preserved if the naturalness in the structure is lost, that preservation should be giver lower weightage. This show the importance of finding the location where the structural information is preserved and the structures are natural. This is called a visual model which gives different weightage to different structural modifications. Finding a visual model is the biggest challenge of formulating an objective metric.

4.2 Parameter Selection

The objective evaluation of tone mapping algorithms motivated us to select two parameters structural similarity and naturalness for the formulation of a good objective quality metric. The most natural preserved structural information should be given high weightage for calculation of final objective quality score. The naturalness map of the image is used as visual model. The naturalness value in a small subregion of an image is considered as a weightage to the structural similarity in that local neighbourhood. Thus highest natural structural changes are given highest weightage.

4.2.1 Structural Similarity

The humans have a good knowledge of most of the natural structures. So even though the subjects didn't have any knowledge about the structure of the input reference image the subjects were able to rate the algorithm output with some amount of accuracy. The FSIM has the highest correlation with the subjective scores which proves the fact that, in the highly rated images high amount of structure is preserved and in low subjective rated images have very low structural preservation. Even if the images have high structural information preserved were given moderate ratings. This proves the fact that in highly structure preserved images, the importance was given to naturalness. We have used SSIM map the map of structural preservation over the entire image.

4.2.2 Naturalness

The SBIQE motivated by human visual system which computes Sparsity coefficients at different locations of an image, and uses these as a measure of naturalness in that local sub region had highest correlation with the subjective scores. This motivates us to use sparsity as a feature to compute the naturalness at a local sub region of an image. This is used as a naturalness map of an image. But the SBIQE wont consider the colour information. By our survey of the Database we found that the images have high amount of unnatural colour artifacts. So a new naturalness map which considers the colour information should be formulated. We proved the assumption of considering both structural information and naturalness while calculating the objective algorithm performance score by using SBIQE map.

4.3 CNN based visual model

The importance of including colour information motivated us to formulate a Convolutional Neural Network (CNN) based visual model to compute the naturality map. Convolutional neural network is a traditional neural network which learns shared convolutional kernels and uses output of these kernels as features in training fully connected neural network. Each convolutional kernel concentrates on different visual details like edges, colour etc. We need a very large number of images to train CNN based network in order to compute these kernel weights, which acts as features to efficiently represent an image. But the image set we have was 420 which was not enough to learn a deep network. Usually in literature this type of problem is addressed by subsampling the image into multiple segments and use each sample as an input image for training which increases the number of training images. This requires an assumption of homogeneous distribution of artifacts introduced. This class of images doesn't have homogeneous spread of distortions which leads us to consider pretrained model. We used pretrained model (IMAGENET) [31] model to compute CNN features and used these features to train the fully connected network for computing naturalness score at each subregion of an image. The accuracy of the CNN model is 62. The correlation of the test image objective naturalness score with the subjective score is 0.62. This can improved. This is more than the naturalness score computed by using SBIQE which was 0.53. This CNN model addressed the problem of including colour information in assessing the naturalness of an image.

4.4 Objective Metric

The objective metric was formulated which uses direct multiplication of local naturalness and structural similarity values to compute an overall algorithm performance map. The average of all the local performance measures is the final output objective performance score for an input image.

Chapter 5

Subjective Evaluation of De-Ghosting algorithms

The De-ghosting algorithms uses multiple images with different exposure values for computing an image with improved visual details. The efficient fusion of these images produces a De-Ghosted image. The assumption to fuse these images together is that the scene is static. The natural scenes will be mostly dynamic. The process of converting all the images to a static scenario taking one of the input LDR images is called as alignment. If the alignment of the dynamic scene is imperfect then there is scope for GHOSTs. The De-Ghosting algorithms first aligns the images and then use fusing . And after fusing the images together algorithms remove the ghosts present in the fused image by using different approaches like rank minimization etc. The De-Ghosting algorithm performance depends on two parameters improvements in the visual contents and the amount of ghosts present in the algorithm output. And the visual details present in the reference image are also very important to the human perception. The algorithm should preserve these details. And the extra contents added should be natural. Even if the amount of details added are very high if the naturalness of the added details is below the acceptable limit of the humans then the algorithm performance is considered to be low. There is a trade off between naturalness and improvement in visual details. The overall algorithm performance by observation we can say that it depends on these two parameters naturalness and improvements in visual details. The De-Ghosting algorithm generation is a very fast growing research area. There is no standard database available to test the performance of different De-Ghosting algorithms. The subjective evaluation of these De-Ghosting algorithms is of utmost importance to be able to capture the state-of-the-art performance of these algorithms and provide a benchmark standard for the research community for the development of new De-Ghosting algorithms.

De-ghosting algorithms will improve the visual details as specified in the example. The Figure. shows that the added details will have various unnatural artifacts like colour dots, blur, saturation, modification of average luminance value, unnatural edges, and ghosts. These class of images have very different types of distortions when compared to the Image Datasets global communities. And the pixel spread of these visible distortions depends on the resolution of the input image. The distortions are visible to humans if the cover a lager area in pixels. So for the perceivable distortions depend upon image resolution. Most of the images available to the communities is 1920X1080. We can do better while considering the natural scenes interms of resolution.

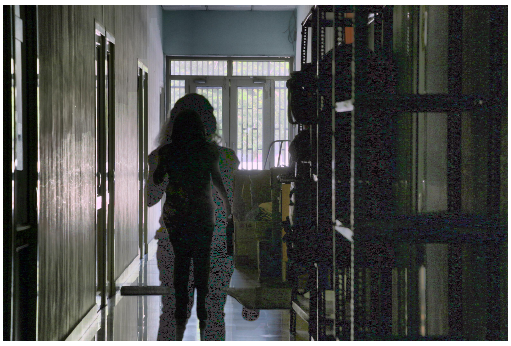


Figure 5.1: De-Ghosted Output

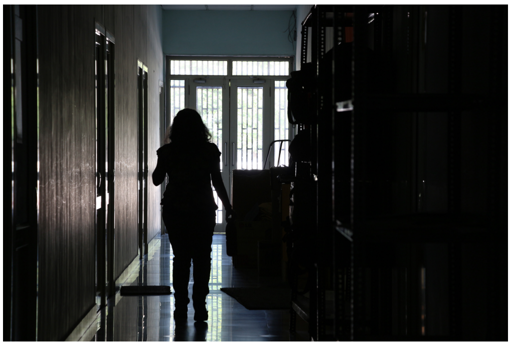


Figure 5.2: De-Ghosting Reference

5.1 De-Ghosting Algorithms

The De-Ghosting algorithm formation is relatively improving area of research. The number of De-Ghosting algorithms publically available for the research communities for academic research purposes is very less. We did an extensive survey to find the recently published state-of-the-art de-Ghosting algorithms for the selection of De-ghosting algorithms and the best parameters for their performance. We selected 6 state-of-the-art De-Ghosting algorithms, [32], [33], [34], [35], Photoshop . We used default parameters after testing with various input parameters for best performance.

5.2 Database

In construction of Database have various challenges like selection of input reference image for alignment, the scene selection which causes the De-Ghosting algorithms to fail. Various types of distortions are included in the algorithm output with varying scene contents like illumination, texture etc. The Database should include most of the possible distortions for this class of images. The selection of scenes should be done in such a way that includes as many distortions as possible. The scenes are selected to create a lot of challenging tasks to the algorithms which in turn creates a lot of visible distortions. Different selection of reference varies the algorithm performance.

5.2.1 Challenges to algorithms

The algorithm output to include most of the distortion. Illumination of the scene varies the distortions included in the output image. If the image is equally illuminated, most of the details captured in the different exposure values will not be much different. In this case the alignment of the LDR stack is simpler. If the illumination of the scene is different at different location each exposure image contains different information. Depending on the exposure value of the camera the camera would either concentrate on highly illuminated region which makes the poorly illuminated region to be completely dark without any visual details or low illumination regions which will capture details in darker regions and will inturn make the bright regions to be saturated. This causes the alignment to be imperfect which involves a lot of distortion types. The texture of the images includes a lot of artifacts in the edges. And the dynamic variation of the scene creates a lot of ghosts. We captured the input LDR scenes with large time difference between the capturing moments. As in the example this time separation make the scene very dynamic. The different capturing scenarios captures different scene contents. Thus the input scenes are captured in such a way to include most of the illumination conditions like equally, partially illuminated regions and very dynamic scenes. The resolution of the images also plays a major role in perceiving visible distortions. We should be able to project these high resolution images on displays in order to perceive these distortion. So the resolution of images was limited by the resolution of display. The mostly available highest resolution was 3840X2160. We captured images to fit these specifications. We included some standard images with resolution of 1920X1080 also which constructing a database which are included in the evaluatiation of most of the standard De-Ghosting algorithms.

5.2.2 Selection of reference

Selection reference effects the output significantly. As the details present in the reference image are very important to the subject, various input LDRs capture different visual information. The image which produces best algorithm output was selected as reference. This was done by manually varying input LDR. This was motivated by the fact that, any algorithm should represent the scene efficiently with most of the scene contents captured. Thus the selection of reference is an important criteria in the algorithm performance. We selected the best scene representation as a metric to evaluate the performance of the algorithm. This survey of input LDR images motivates us to formulate an objective algorithm for the selection of reference image.

5.3 Experimental Setup and GUI

We used 4K LG tv with resolution of 3840X2160 for display purposes. The subjective evaluation is very time consuming and hectic task. The subject was provided a Keyboard and GUI to make the subjective evaluation easier. The subject viewed the display from a distance of 2m according the optimal viewing distance of the LG 4k TV. The GUI was as shown in the Figure 5.3 below.

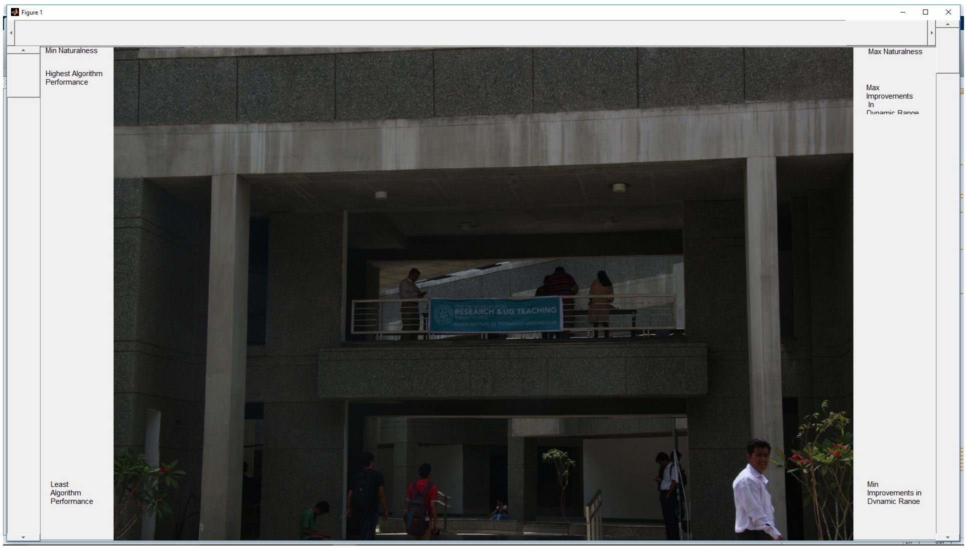


Figure 1
Min Naturalness
Highest Algorithm Performance
Least Algorithm Performance
Max Naturalness
Max Improvements In Dynamic Range
Min Improvements in Dynamic Range
RESEARCH & UG TEACHING

Figure 5.3: De-Ghosting GUI

5.4 Subjective Evaluation

The subjective evaluation was done using Double Stimulus Continuous Quality evaluation (DSCQE) method which was on of the standard methods. We made the subjective evaluation to be continuous because of the fact the subjects are going to rate the algorithms relatively. So to provide a handle to the subject to rate the algorithms relatively with most accuracy a continuous slider was used. The continuous slider accept accepts the inputs with an accuracy of 4 decimal places. According to standards the highest possible precision with which a subject can evaluate is upto two decimal places. The input of slider was rounded off to two decimal places in order to match the performance of the subjects. The maximum performance score was 10, and minimum performance score was 0. The subject was asked to enter two parameters naturalness of the algorithm output, improvements in the visual details and overall performance score for the algorithm. The study was done with 25 subjects validated by using outlier removal method mentioned in . This is the standard number of images used according to standards of DSCQE. The time duration is another important metric while doing a subjective study. Because with larger time slots the subjects tend to deviate from their natural decision. According to literature (DSCQE) the allowable time duration for a single slot quality evaluation is 30 to 60 minutes. Typically it was found the the evaluation took at an average 45 minutes. So the the evaluation was done in on continuous time slot. We used one test scene to make the subject familiar with the GUI. After training the subjects were allowed to do the subjective evaluation continuously for one time slots. All the observations were captured in similar laboratory conditions with same brightness and illumination throughout the study. The study was full reference. The subject was show 120 algorithm outputs and 120 reference images. The ordering of the algorithm output was randomised in order to capture the mean opinion, as the the subjective evaluation was capturing relative algorithm performance. Each time 6 randomised algorithm outputs were shown to the subject before moving on the next scene.

Chapter 6

Objective evaluation of De-Ghosting Algorithms

The objective evaluation of these algorithms is necessary because of the fact that always the subjective evaluation is very time consuming and the accuracy of the subjects is very low. And an extensive survey was done in order to find a suitable objective quality metric to evaluate this class of images. It was found the specific algorithm to objectively assess the quality of the images was not present in literature. So the best match algorithms were selected to objectively evaluate the performance of the algorithms.

6.1 Objective Metric Selection

The selection type of objective metric was very important. The subjective study done was full reference subjective study. The evaluation of the performance of the algorithms needs a reference image to compare the algorithm output to find the performance. But the visual details of the subjects are very much modified as shown in the example figure. So the existing state-of-the-art full reference metrics fail to compute the objective quality score of the algorithms. Because the visual details are modified which makes the structure modified. Another important metric in assessing the performance of the algorithms is naturalness of the algorithm output. The no reference objective quality metrics consider this naturalness as the main parameter to evaluate the quality of the images. So we selected state of the best performing algorithms no reference objective quality metrics QAC [24], NIQE [26] and SBIQE [25] for objective evaluation.

6.2 Results and Discussion

Subjective Score	QAC	NIQE	SBIQE
Naturalness	0.47	0.42	0.48
Overall Performance	0.35	0.33	0.27

Table 6.1: Correlation of Subjective Scores with Objective Scores.

The correlation of the objective scores with subjective scores is reported in the Table.6.1. This shows that the performance the existing objective quality metrics is very poor. The no reference quality metric QAC performs better with correlation of 0.35. And the correlation of SBIQE scores with the naturalness is very high. It shows that the sparsity based objective evaluation best captures the naturalness of a De-Ghosted image. There is no existing algorithm to compute the improvements in the visual details. The overall performance score not only depends on the naturalness but it also depends on the improvements in visual details which is improvements in dynamic range. There is necessity of finding a metric to calculate the dynamic range score and then combine it the naturalness score to compute the overall performance of the algorithm.

Chapter 7

Formulation of Objective Metric for the evaluation of De-Ghosting Algorithms

From the correlations with the subjective scores it is evident that there is necessity of formulating an objective quality assessment algorithm to find the performance of the algorithms. And the parameters to consider are naturalness and improvements in dynamic range and combination of these scores to compute the overall performance score.

7.1 Challenges

The artifacts present in the images are very different from the existing artifacts. A lot of ghosts are included in the algorithm output which are very new type of distortion. So the existing naturalness metrics fail in capturing the naturalness of the algorithm output. This is evident by seeing the correlation of the naturalness subjective scores with the existing objective algorithms is moderate. Showing the new distortions are included. And the objective algorithm should be perceptually motivated in order to best match the subjective scores and the end user of the algorithms are humans. The colour information is mostly modified so the algorithm should consider colour information while evaluating the performance. And the algorithm should be ful reference. The distortion are not distributed homogeneously over the entire image. The current best performing algorithm SBIQE is statistical not learning based which inspires us to use sparsity as one of the features in formulating a perceptual learning based algorithm. And other features should be included in order to include colour information and ghost artifacts.

7.2 Relative Naturalness

As we have seen the naturalness is very important metric to evaluate the performance of the algorithm. The naturalness computed for an image is no reference. Even the best performing algorithms might include distortions if the distortions are originally present in the input LDR stack. So the distortions can be present in the algorithm output if at a particular local region has distortions in the original input image. This is the relative naturalness measure of an algorithm output with reference to input. The amount of distortion present in the inputs to the algorithm can be computed by considering either only the reference or the entire LDR stack. The details present in the reference are very important and reference image contains most of the information of the scene so the relative naturalness was found with respect to the reference. The correlation of the relative naturalness found by using only sparsity as the feature was 0.52 which was greater that naturalness of the output alone which was 0.48. The small improvement in naturalness is because most of the input reference images are highly natural not modifying the relative naturalness score much. Only some of the input LDR images have low naturalness.

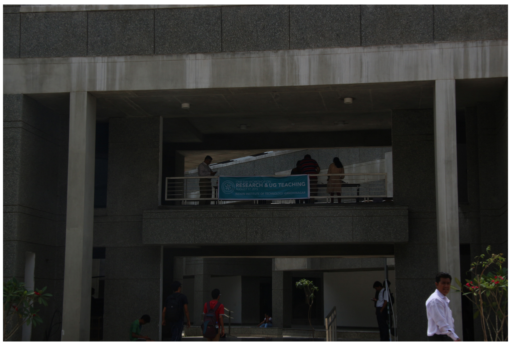


Figure 7.1: Deghosting Algorithm input1

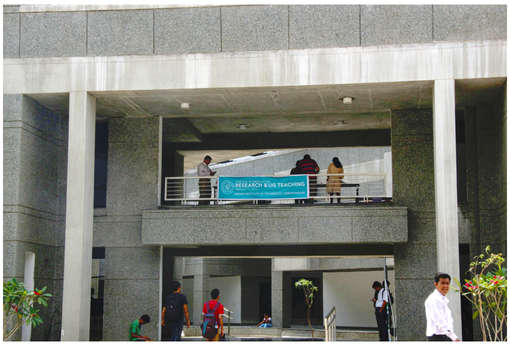


Figure 7.2: Deghosting algorithm outpu1 with RDR of 1.15

7.3 Relative Dynamic range

Another important metric in evaluating the performance of the De-Ghosting algorithms is the Dynamic Range of the algorithm output with reference to the input LDR. Again only the input reference image is used for computing the relative Dynamic range measure. The dynamic range of the image can be used as a feature to capture the improved visual details. The relative dynamic range has a correlation of 0.58 with subjective visual detail improvement scores. Motivating us to use dynamic range as a feature in computing the visual detail improvement scores. But the correlation is still moderate. Because the naturalness of the improvements also affects the subjective perception. So

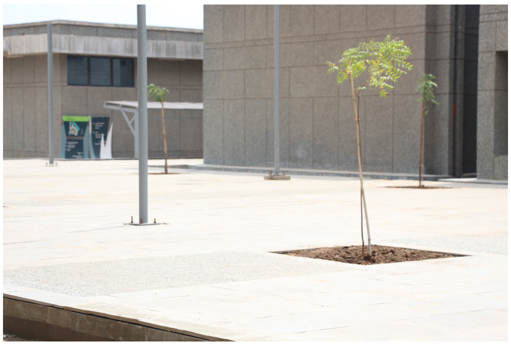


Figure 7.3: Deghosting Algorithm input2

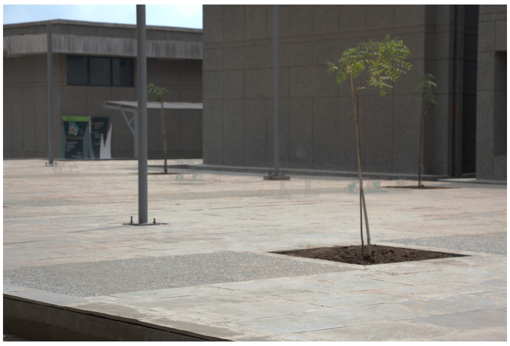


Figure 7.4: Deghosting algorithm output2 with RDR of 1. 05

even the naturalness is involved in computing the dynamic range. It is not independent. So the dynamic range feature can be used and a visual model should be used to compute effective dynamic range improvements to improve the correlation with the subjective scores. The overall dynamic range of the image is computed by averaging local dynamic ranges computed by using local windows of various sizes. The optimal window size depends on the image resolution. For the resolution of image used an empirical value of window size of 27X27 was used. The dynamic range comparison can seen by observing the figures Fig.7.1, Fig.7.2, Fig.7.3, Fig.7.4. The Fig.7.1, Fig.7.3 the images are the inputs to the De-Ghosting algorithm Photoshop, and Fig.7.2, Fig.7.4 are the outputs of the algorithm. The visible improvements in Fig.7.1, are more when compared to Fig.7.3 that is evident by observing the Dynamic range scores stated below the figures.

Figure 7.5: MSCN of 1.4

7.4 Differenciation between Input LDR images using Mscn Coefficients

Another main task for the objective algorithm to find the performance the selection of best reference image. This was done manually for the subjective study. But always the subjective input will not be available. So we should be able to differentiate between the input LDRs and with help of subjective study we can learn using this feature to find the best performing reference algorithmically. The variance of the histogram of MSCN coefficients can be used as a feature to differentiate between different LDR inputs. The window size of 13X13 was empirically chosen to find MSCN coefficients. The comparison of MSCN coefficients can be done by observing the figures Fig.7.5, Fig.7.6, Fig.7.7. The mscn coefficient of the moderately illuminated image Fig.7.6 has the highest value. And the least illuminated, highest illuminated images have moderate value of variance of histogram of MSCN coefficients.

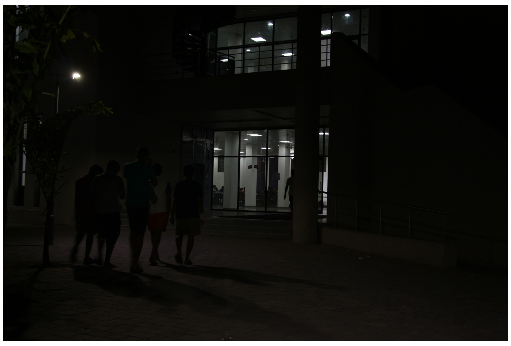


Figure 7.6: MSCN of 2.4

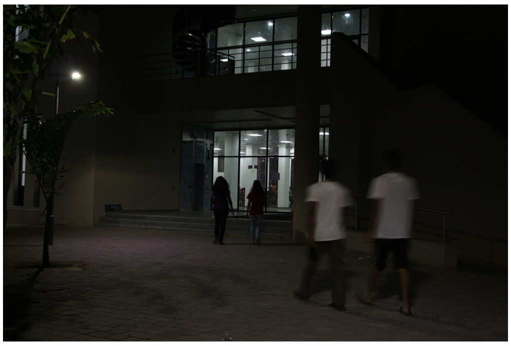


Figure 7.7: MSCN of 1. 05

7.5 Learing Overall Performance score of Algorithms

The relative naturalness and relative dynamic range are two main parameters in computing the overall algorithm performance. The combination of these can be nonlinear. So we used MLP network to learn the relation between the relative naturalness, relative dynamic range and overall performance score. The network parameters are chosen empirically to have maximum accuracy of

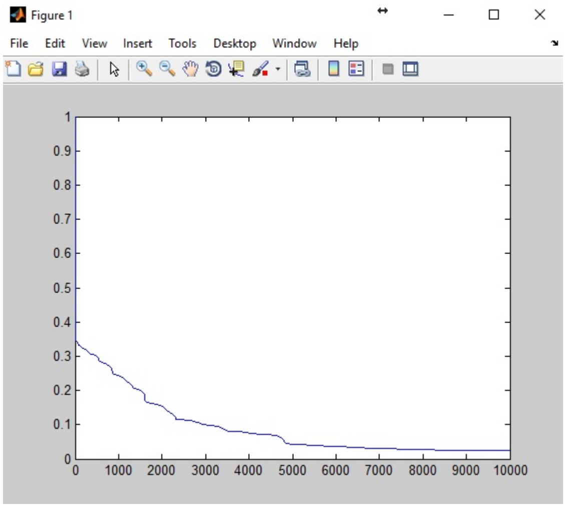


Figure 7.8: Error Vs Iterations

0.96. The network implemented has 5000 nodes in each hidden layer and 2 hidden layers. One fifth of the Database are used for validating and others for training. Accuracy of 0.96 was achieved which shows that the overall performance score is a combination of relative naturalness score and relative improvements score. The Figure Fig.7.8shows the error curve for 25000 iterations. We used 25000 iterations while learning the overall algorithm performance score.

7.6 Overall Algorithm performance Score

The overall algorithm performance score is the average of local relative naturalness and relative dynamic range. The learnt relation was used in combining the local scores also. Then an average score was found.

Chapter 8

Future Work

For including the colour information learning based method is needed. But the learning based method needs a large Dataset or pretrained model. Pretrained model for this class of images is not present. The resolution of input images is very large. An effective way to subsample the input image should be found in order to represent the image. And the pretrained models can be used to find the local features and then combine all the features together to find bigger feature set and compute an overall performance score. Instead of average an efficient way of computing an overall performance score can be formulated.

References

[1] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski. High dynamic range imaging: acquisition, display, and image-based lighting. Morgan Kaufmann, 2010.

[2] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers. Advanced High Dynamic Range Imaging: Theory and Practice. AK Peters (CRC Press), Natick, MA, USA, 2011.

[3] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Dynamic range independent image quality assessment. In ACM Transactions on Graphics (TOG), volume 27. ACM, 2008 69.

[4] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang. High Dynamic Range Image Compression by Optimizing Tone Mapped Image Quality Index. *Image Processing, IEEE Transactions on* PP, (2015) 1–1.

[5] M. Ashikhmin. A tone mapping algorithm for high contrast images. In Proceedings of the 13th Eurographics workshop on Rendering. Eurographics Association, 2002 145–156.

[6] F. Banterle, A. Artusi, E. Sikudova, T. Bashford-Rogers, P. Ledda, M. Bloj, and A. Chalmers. Dynamic range compression by differential zone mapping based on psychophysical experiments. In Proceedings of the ACM Symposium on Applied Perception. ACM, 2012 39–46.

[7] F. Drago, K. Myszkowski, T. Annen, and N. Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In Computer Graphics Forum, volume 22. Wiley Online Library, 2003 419–426.

[8] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM transactions on graphics (TOG)* 21, (2002) 257–266.

[9] R. Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. In ACM Transactions on Graphics (TOG), volume 21. ACM, 2002 249–256.

[10] P. Irawan, J. A. Ferwerda, and S. R. Marschner. Perceptually Based Tone Mapping of High Dynamic Range Image Streams. In Rendering Techniques. 2005 231–242.

[11] M. H. Kim, T. Weyrich, and J. Kautz. Modeling human color perception under extended luminance levels. In ACM Transactions on Graphics (TOG), volume 28. ACM, 2009 27.

[12] G. Krawczyk, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Lightness perception inspired tone mapping. In Proceedings of the 1st Symposium on Applied perception in graphics and visualization. ACM, 2004 172–172.

[13] J. Kuang, G. M. Johnson, and M. D. Fairchild. iCAM06: A refined image appearance model for HDR image rendering. *Journal of Visual Communication and Image Representation* 18, (2007) 406–414.

[14] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. In ACM Transactions on Graphics (TOG), volume 27. ACM, 2008 67.

[15] E. Reinhard and K. Devlin. Dynamic range reduction inspired by photoreceptor physiology. *Visualization and Computer Graphics, IEEE Transactions on* 11, (2005) 13–24.

[16] C. Schlick. Quantization techniques for visualization of high dynamic range pictures. In Photorealistic Rendering Techniques, 7–20. Springer, 1995.

[17] G. J. Ward. The RADIANCE lighting simulation and rendering system. In Proceedings of the 21st annual conference on Computer graphics and interactive techniques. ACM, 1994 459–472.

[18] G. W. Larson, H. Rushmeier, and C. Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *Visualization and Computer Graphics, IEEE Transactions on* 3, (1997) 291–306.

[19] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg. Time-dependent visual adaptation for fast realistic image display. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 2000 47–54.

[20] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. In ACM Transactions on Graphics (TOG), volume 21. ACM, 2002 267–276.

[21] G. Ward. A contrast-based scalefactor for luminance display. *Graphics gems IV* 415–421.

[22] J. Tumblin and H. Rushmeier. Tone reproduction for realistic images. *Computer Graphics and Applications, IEEE* 13, (1993) 42–48.

[23] I. R. Assembly. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, 2003.

[24] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013 995–1002.

[25] K. Priya and S. S. Channappayya. A novel sparsity-inspired blind image quality assessment algorithm. In Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on. IEEE, 2014 984–988.

[26] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *Signal Processing Letters, IEEE* 20, (2013) 209–212.

[27] L. Zhang, D. Zhang, and X. Mou. FSIM: a feature similarity index for image quality assessment. *Image Processing, IEEE Transactions on* 20, (2011) 2378–2386.

[28] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel. Predicting visible differences in high dynamic range images: model and its calibration. In Electronic Imaging 2005. International Society for Optics and Photonics, 2005 204–214.

[29] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In ACM Transactions on Graphics (TOG), volume 30. ACM, 2011 40.

[30] M. A. Krishna, S. S. Chandra, S. S. Channappayya, and S. Raman. A subjective and objective quality assessment of tone-mapped images. In 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP). 2015 443–447.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds., Advances in Neural Information Processing Systems 25, 1097–1105. Curran Associates, Inc., 2012.

[32] C. Lee, Y. Li, and V. Monga. Ghost-Free High Dynamic Range Imaging via Rank Minimization. *IEEE Signal Processing Letters* 21, (2014) 1045–1049.

[33] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman. Robust Patch-based Hdr Reconstruction of Dynamic Scenes. *ACM Trans. Graph.* 31, (2012) 203:1–203:11.

[34] J. Hu, O. Gallo, K. Pulli, and X. Sun. HDR Deghosting: How to Deal with Saturation? In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. 2013 1163–1170.

[35] S. Raman and S. Chaudhuri. Reconstruction of High Contrast Images for Dynamic Scenes. *Vis. Comput.* 27, (2011) 1099–1114.