SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT OF ANCIENT DEGRADED DOCUMENTS

by

Atena SHAHKOLAEI

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, SEPTEMBER 27, 2019

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Mohamed Cheriet, Thesis Supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Luc Duong, President of the Board of Examiners
Department of Software Engineering and IT, École de technologie supérieure

Mr. Stéphane Coulombe, Member of the jury
Department of Software Engineering and IT, École de technologie supérieure

Mr. Ching Y. Suen, External Independent Examiner
Department of computer science and software engineering, Concordia University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON NOVEMBER 11TH, 2019

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

This thesis becomes a reality with the kind supports and help of many individuals. I would like to extend my sincere thanks to all of them.

I find myself fortunate to be blessed with very supporting adviser Prof. Mohamed Cheriet, who have been encouraging me all these research years. Thanks a lot for all the essential advice and permanent guidance that I received from him, for showing the path of research, for creating an enjoyable working environment, and for imparting his knowledge and experience in the whole duration of my Ph.D.

Besides my advisor, I want to thank the rest of my thesis committee: Prof. Stéphane Coulombe, Prof. Luc Duong, and Prof. Ching Y. Suen for reviewing my dissertation and providing me with valuable comments.

I would like to thank Prof. Azeddine Beghdadi for his collaboration in my thesis. I must also thank all my friends in the Synchromedia Laboratory.

The most significant acknowledgment belongs to my darling husband Hossein for his synergism in my scientific and personal life. Existence of this thesis would not be possible without him. Many thanks go to my lovely son, Atila, who is my inspiration to achieve greatness.

Last but certainly not least, I am very thankful for my dear family, especially my sacred mother, who supported me for my progress and immensely.

In memory of my father

ÉVALUATION SUBJECTIVE ET OBJECTIVE DE LA QUALITÉ DE DOCUMENTS ANCIENS DÉGRADÉS

Atena SHAHKOLAEI

RÉSUMÉ

L'archivage, la restauration et l'analyse d'anciens manuscrits ont considérablement augmenté au cours des dernières décennies. Habituellement, ces documents sont physiquement dégradés à cause du vieillissement et d'une manipulation incorrecte. Ils ne peuvent pas non plus être traités manuellement car un grand nombre de ces documents existent dans les bibliothèques et les archives du monde entier. Par conséquent, des méthodologies automatiques sont nécessaires pour préserver et traiter leur contenu. Ces documents sont généralement traités à travers leurs images. Le traitement des images de documents dégradés est une tâche difficile, principalement en raison des dégradations physiques existantes. Bien qu'il puisse être très difficile de localiser et d'éliminer avec précision ces distorsions, il est possible d'analyser la gravité et le type de ces distorsions. Cette analyse fournit des informations utiles sur le type et la gravité des dégradations dans un certain nombre d'applications. Les principales contributions de cette thèse sont de proposer des modèles d'évaluation objective de la condition physique des images de documents et de classer leurs dégradations. Dans cette thèse, trois ensembles de données d'images de documents dégradés ainsi que les évaluations subjectives de chaque image sont développés. En outre, trois mesures d'évaluation de la qualité d'image des documents sans référence (NR-DIQA) sont proposées. Enfin, nous proposons un modèle de classification de dégradation afin d'identifier les types de distorsion courants dans les anciennes images de document.

Essentiellement, les métriques d'évaluation de qualité d'image sans référence (NR-IQA) existantes ne sont pas conçues pour évaluer les distorsions physiques des documents. Dans la première contribution, nous proposons le premier ensemble de données d'images de documents dégradés ainsi que les scores d'opinion humaine pour chaque image de document. Cet ensemble de données est introduit pour évaluer la qualité des images de documents historiques. Nous proposons également une métrique objective NR-DIQA basée sur les statistiques des coefficients de moyenne normalisée avec contraste soustrait (MSCN) calculés à partir des couches segmentées de chaque image de document. La segmentation en quatre couches d'avant-plan et d'arrière-plan est réalisée à partir d'une analyse des filtres de log-Gabor. Cette segmentation est basée sur l'hypothèse que la sensibilité du système visuel humain (HVS) est différente aux emplacements texte et non texte. Les résultats expérimentaux montrent que la métrique proposée a des performances comparables ou meilleures que les métriques de pointe, alors qu'elle a une complexité modérée.

L'identification de la dégradation et l'évaluation de la qualité peuvent se compléter pour fournir des informations sur le type et la gravité des dégradations dans les images de document. Par conséquent, nous avons présenté dans la deuxième contribution une base de données d'images de documents historiques multi-distorsions pouvant être utilisée pour la recherche sur l'évaluation de la qualité des documents dégradés ainsi que sur la classification de la dégradation. Le jeu de données développé contient des images de documents historiques classées en quatre catégories en fonction de leur type de distorsion, à savoir la translucidité du papier, la teinture, les annotations du lecteur et les trous usés. Une métrique NR-DIQA efficace est ensuite proposée sur la base de trois ensembles d'entités d'images spatiales et fréquentielles extraites de deux couches de texte et de non-texte. En outre, ces caractéristiques sont utilisées pour estimer la probabilité des quatre distorsions physiques susmentionnées pour la première fois dans la littérature. Les modèles proposés d'évaluation de la qualité et de classification de la dégradation offrent des performances très prometteuses.

Enfin, nous développons dans la troisième contribution un ensemble de données et une mesure d'évaluation de la qualité pour les images de documents graphiques dégradés (DMD). Ce type d'images dégradées contient des informations textuelles et picturales. Le jeu de données DMD introduit est le premier jeu de données de sa catégorie qui fournit également des évaluations humaines. Nous proposons également une nouvelle métrique sans référence afin d'évaluer la qualité des images DMD dans l'ensemble de données développé. La métrique proposée est basée sur l'extraction de plusieurs entités statistiques à partir de trois couches de texte, non textuelles et graphiques. La segmentation est basée sur la saillance des couleurs en supposant que les parties illustrées sont colorées. Il suit également le HVS qui attribue des poids différents à chaque couche. Les résultats expérimentaux valident l'efficacité de la stratégie proposée NR-DIQA pour les images DMD.

Mots-clés: Perception de la vision humaine, Évaluation de la qualité d'image, Pas de référence, Note moyenne d'opinion, Contraste moyen soustrait normalisé, Phase locale, Approche de saillance, Classification de la dégradation, Machine à vecteurs de support, Documents médiévaux dégradés.

SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT OF ANCIENT DEGRADED DOCUMENTS

Atena SHAHKOLAEI

ABSTRACT

Archiving, restoration and analysis of damaged manuscripts have been largely increased in recent decades. Usually, these documents are physically degraded because of aging and improper handing. They also cannot be processed manually because a massive volume of these documents exist in libraries and archives around the world. Therefore, automatic methodologies are needed to preserve and to process their content. These documents are usually processed through their images. Degraded document image processing is a difficult task mainly because of the existing physical degradations. While it can be very difficult to accurately locate and remove such distortions, analyzing the severity and type(s) of these distortions is feasible. This analysis provides useful information on the type and severity of degradations with a number of applications. The main contributions of this thesis are to propose models for objectively assessing the physical condition of document images and to classify their degradations. In this thesis, three datasets of degraded document images along with the subjective ratings for each image are developed. In addition, three no-reference document image quality assessment (NR-DIQA) metrics are proposed for historical and medieval document images. It should be mentioned that degraded medieval document images are a subset of the historical document images and may contain both graphical and textual content. Finally, we propose a degradation classification model in order to identify common distortion types in old document images.

Essentially, existing no reference image quality assessment (NR-IQA) metrics are not designed to assess physical document distortions. In the first contribution, we propose the first dataset of degraded document images along with the human opinion scores for each document image. This dataset is introduced to evaluate the quality of historical document images. We also propose an objective NR-DIQA metric based on the statistics of the mean subtracted contrast normalized (MSCN) coefficients computed from segmented layers of each document image. The segmentation into four layers of foreground and background is done based on an analysis of the log-Gabor filters. This segmentation is based on the assumption that the sensitivity of the human visual system (HVS) is different at the locations of text and non-text. Experimental results show that the proposed metric has comparable or better performance than the state-of-the-art metrics, while it has a moderate complexity.

Degradation identification and quality assessment can complement each other to provide information on both type and severity of degradations in document images. Therefore, we introduced, in the second contribution, a multi-distortion historical document image database that can be used for the research on quality assessment of degraded documents as well as degradation classification. The developed dataset contains historical document images which are classified into four categories based on their distortion types, namely, paper translucency, stain, readers' annotations, and worn holes. An efficient NR-DIQA metric is then proposed based on three sets of spatial and frequency image features extracted from two layers of text and non-text. In addition, these features are used to estimate the probability of the four aforementioned physical distortions for the first time in the literature. Both proposed quality assessment and degradation classification models deliver a very promising performance.

Finally, we develop in the third contribution a dataset and a quality assessment metric for degraded medieval document (DMD) images. This type of degraded images contains both textual and pictorial information. The introduced DMD dataset is the first dataset in its category that also provides human ratings. Also, we propose a new no-reference metric in order to evaluate the quality of DMD images in the developed dataset. The proposed metric is based on the extraction of several statistical features from three layers of text, non-text, and graphics. The segmentation is based on color saliency with assumption that pictorial parts are colorful. It also follows HVS that gives different weights to each layer. The experimental results validate the effectiveness of the proposed NR-DIQA strategy for DMD images.

Keywords: Human vision perception, Image quality assessment, No-reference, Mean opinion score, Mean subtracted contrast normalized, Local phase, Saliency approach, Degradation classification, Support vector machine, Degraded medieval documents.

TABLE OF CONTENTS

Page

english

INTRODUCTION 1
0.1 Problem Statement (PS) 3
0.1.1 PS1. No dataset 3
0.1.2 PS2. Physical degradations 4
0.1.3 PS3. Limited research 4
0.1.4 PS4. No quality assessment metric for historical documents 5
0.1.5 PS5. No degradation identification model for HDIs 6
0.2 Research questions 6
0.2.1 Research Question (RQ1) 6
0.2.2 Research Question (RQ2) 6
0.2.3 Research Question (RQ3) 7
0.3 Contributions 7
0.4 Structure of the thesis 9

CHAPTER 1 LITERATURE REVIEW 11
1.1 Image quality assessment 11
1.1.1 Subjective image quality assessment 11
1.1.1.1 Direct scaling methods 12
1.1.1.2 Indirect scaling methods 13
1.1.2 Subjective document image quality assessment and datasets 14
1.1.3 Objective image quality assessment 15
1.1.3.1 FR (Full Reference) 16
1.1.3.2 RR (Reduced Reference) 17
1.1.3.3 NR (No-Reference) 18
1.1.4 Objective document image quality assessment 20
1.1.4.1 OCR accuracy 21
1.1.4.2 Human perception 22
1.1.5 Discussion 23
1.2 Distortion identification algorithm 23
1.2.1 Natural images 23
1.2.2 Ancient images 25
1.2.3 Discussion 25
1.3 Evaluation measures for objective quality assessment 25
1.3.1 Pearson's Linear Correlation Coefficient (PLCC) 26
1.3.2 Spearman's Rank Order Correlation Coefficient (SROCC) 26
1.3.3 Kendall's Rank order Correlation Coefficient (KRCC) 26
1.3.4 Root Mean Square Error (RMSE) 27

CHAPTER 2 OBJECTIVES AND GENERAL METHODOLOGY 29

2.1 Research objectives 29
2.1.1 Objective O1: Subjective and objective quality assessment of historical document images based on human judgments (PS1-PS4 and RQ1) 29
2.1.2 Objective O2: Subjective and objective quality assessment as well as degradation identification for physical distortions in damaged manuscripts (PS1-PS5 and RQ2) 30
2.1.3 Objective O3: Propose a new dataset and a quality assessment metric for degraded medieval document images based on a saliency approach (PS1-PS4 and RQ3) 30
2.2 General methodology 31
2.2.1 Methodology M1: First dataset and efficient blind IQA metric for degraded documents 31
2.2.2 Methodology M2: A new blind quality assessment metric, dataset and degradation identification algorithm 32
2.2.3 Methodology M3: Subjective and objective quality assessment metric for degraded medieval documents 32

CHAPTER 3 ARTICLE I- SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT OF DEGRADED DOCUMENT IMAGES 35
3.1 Introduction 36
3.2 VDIQA dataset 41
3.2.1 Subjective evaluation 42
3.3 Proposed VDIQA metric 46
3.3.1 Segmentation based on log-Gabor filter 47
3.3.2 Spatial domain document statistics 49
3.3.2.1 Symmetric generalized Gaussian distribution 52
3.3.2.2 Asymmetric generalized Gaussian distribution 53
3.4 Experimental results and discussion 54
3.4.1 Performance comparison 54
3.4.2 Computational Analysis 56
3.5 Applications 57
3.6 Conclusion 58

CHAPTER 4 ARTICLE II- MHDID: A MULTI-DISTORTION HISTORICAL DOCUMENT IMAGE DATABASE 59
4.1 Introduction 60
4.2 MHDID dataset description 62
4.2.1 Selection of distorted images 62
4.2.2 Subjective image quality evaluation 63
4.2.2.1 Direct scaling methods 63
4.2.2.2 Indirect scaling methods 63
4.2.3 Test environment and data collection 65
4.3 Dataset analysis 68

4.4 Conclusion 69

CHAPTER 5 ARTICLE III- BLIND QUALITY ASSESSMENT METRIC AND DEGRADATION CLASSIFICATION FOR DEGRADED DOCUMENT IMAGES 71
5.1 Introduction 72
5.2 Proposed Document Image Quality Measure (DIQM) 76
5.2.1 Text/non-text image segmentation 77
5.2.2 Document statistical analysis based on the MSCN coefficients 80
5.2.3 Local phase information of RGB channels 82
5.2.4 Symmetric and asymmetric generalized Gaussian distribution 85
5.3 Degradation identification of degraded document images 87
5.4 Experimental results and discussion 88
5.4.1 Performance comparison of the proposed metric 89
5.4.2 Performance comparison of degradation classification model 90
5.4.3 Computational load analysis 92
5.5 Conclusion 93

CHAPTER 6 ARTICLE IV- DEGRADED MEDIEVAL DOCUMENTS: DATASET AND QUALITY ASSESSMENT BASED ON A SALIENCY APPROACH 95
6.1 Introduction 96
6.2 DMDD dataset 98
6.2.1 Selection of degraded images 98
6.2.2 Subjective evaluation 98
6.2.3 Data processing 99
6.3 Proposed metric (BQAD) 100
6.3.1 Segmentation 100
6.3.1.1 Color/graphical layer segmentation 101
6.3.1.2 Background and text segmentation 104
6.3.2 MSCN coefficients 104
6.3.3 Color gradient statistics 106
6.3.4 Feature extraction 106
6.4 Experimental results 107
6.4.1 Computational complexity 109
6.5 Conclusion 110

CHAPTER 7 CRITICAL DISCUSSION OF SOME CONCEPTS OF THE THESIS 113
7.1 Subjective and objective quality assessment of degraded document images (RQ1) 113
7.2 Blind quality assessment metric and degradation classification for degraded document images (RQ2) 114

7.3 Subjective and objective quality assessment of DMD images based on a saliency approach (RQ3) 114

CONCLUSION AND RECOMMENDATIONS 115
8.1 General conclusion and future work 115
8.2 Summary of contributions 118
8.3 Articles in peer-reviewed journals and conferences 119
8.4 Other scientific activities 120

BIBLIOGRAPHY 121

LIST OF TABLES

Page

english

Table 0.1 Selected degradation types for modern and ancient documents. 4

Table 0.2 All contributions in the three research works 9

Table 3.1 Performance comparison of the proposed metric (VDQAM) and six blind image quality assessment metrics on the developed dataset. Note that top two metrics are highlighted. 55

Table 3.2 Performance of the individual layers when considered as separate quality assessment models. 56

Table 3.3 Comparison of six NR-IQA metrics in terms of run time and number of features. 57

Table 4.1 Comparison between VDIQA and MHDID datasets. MTD stands for Multiple Types of Distortions. 66

Table 5.1 Performance comparison of the proposed metric (MDQM) and six blind image quality assessment metrics on the VDIQA and MHDID datasets for 20% train and 80% test. Note the metrics that are specifically designed to assess the physical noises. 90

Table 5.2 Performance comparison of the proposed metric (MDQM) and six blind image quality assessment metrics on the VDIQA and MHDID datasets for 50% train and 50% test. Note the metrics that are specifically designed to assess the physical noises. 90

Table 5.3 Performance comparison of the proposed metric (MDQM) and six blind image quality assessment metrics on the VDIQA and MHDID datasets for 80% train and 20% test. Note the metrics that are specifically designed to assess the physical noises. 91

Table 5.4 The comparison of F-measure values for different degradation types in MHDID dataset for the seven NR-IQA metrics. We highlight the top two best performance metrics with boldface. Values are in percent. 91

Table 5.5 Accuracy comparison between six NR-IQA metrics and the proposed metric. Note that top two metrics are highlighted. Values are in percent. 92

Table 5.6 Comparison of run-time (milliseconds/image) and the number of features for seven NR-IQA metrics. 93

Table 6.1 Detailed information about the DMDD dataset. MTD stands for Multiple Types of Distortions. 100

Table 6.2 Performance comparison of the BQAD metric with six IQA algorithms on the DMDD dataset for three cases: 20% train and 80% test, 50% train and 50% test, 80% train and 20% test. 108

Table 6.3 Correlation between three segmented layers of color (L_1), text (L_2) and non-text (L_3) with the human judgments. 109

LIST OF FIGURES

Page

english

Figure 0.1 Examples of ancient document images with different types of degradation. 5

Figure 1.1 Three different objective image quality assessment models. From left to right, full-reference IQA model, reduced-reference IQA model, and no-reference IQA model. 16

Figure 1.2 The flowchart of FR-IQA models. 16

Figure 1.3 Overview on document IQA models 21

Figure 3.1 Screenshot of the interface used for subjective evaluation. 40

Figure 3.2 Two examples of document images in the VDIQA dataset that contain no text. 43

Figure 3.3 Two examples of document images in the VDIQA dataset that contain side-notes and main-body text. 43

Figure 3.4 Six examples of the degraded document images in the developed dataset with their MOS values. Higher MOS values indicate better-quality images. 44

Figure 3.5 Histogram of MOS values in the developed dataset. Lower MOS values indicate lower-quality document images. 45

Figure 3.6 A scatter plot of spatial information (SI) against colorfulness (CF) for 177 document images in the developed dataset. 45

Figure 3.7 Illustration of the proposed metric. 47

Figure 3.8 An example of the proposed document image segmentation method into four layers of text and non-text: (a) original image, (b) possible text pixels, (c) possible degradations far from the text, (d) possible degradations close to the text, and (e) possible non-degraded background pixels. 49

Figure 3.9 Three document images from the developed dataset along with their intensity histogram, and MSCN coefficients histogram. 50

Figure 3.10 Histograms of MSCN coefficients for (a) possible text pixels, (b) possible degraded pixels far from the text, (c) possible degraded pixels close to the text, and (d) possible non-degraded background pixels.......... 51

Figure 3.11 Histograms of MSCN coefficients for four layers of the three document images.......... 52

Figure 3.12 Scatter plots of quality scores against the subjective MOS on the VDIQA dataset for the proposed model VDQAM. These scatters are shown based on the median value of 1000 times train-test for three cases: 20% train 80% test, 50% train 50% test, and 80% train 20% test.......... 55

Figure 4.1 Some representative samples of degraded document images selected from different degradation categories in the MHDID dataset: paper translucency, stains, readers' annotations and worn holes. These degradations are shown from the first row to the forth row, respectively.......... 64

Figure 4.2 Screenshot of the interface used for subjective evaluation. 66

Figure 4.3 Histogram of MOS values in the developed dataset.......... 67

Figure 4.4 A scatter plot of spatial information (SI) against colorfulness (CF) for 335 degraded document images in the MHDID dataset.......... 69

Figure 5.1 The framework of the proposed blind MDQM metric.......... 77

Figure 5.2 Illustrations of the proposed document image segmentation method and the associated MCSN coefficients histograms of text and non-text classes. Foreground and background pixels of two degraded images are shown in (b) and (d) images with white and black color, respectively. In (e) and (f) MCSN histograms where the blue plot is related to the text regions (L_1), while the red plot is related to the non-text regions (L_2). 79

Figure 5.3 Histograms of intensity, MSCN and MSCN of gradient magnitude for three degraded document images with different MOS values, selected from the MHDID dataset.......... 82

Figure 5.4 Histograms of the local phase of two degraded document images ((a)-(b)). The histograms in the second row indicate the distribution of local phase which is computed directly from gray scale version of images ((c)-(d)). The histograms in the third row correspond to the distribution of local phase on the RGB channels separately ((e)-(f)). The histograms of the mean of local phase of R, G and B channels are shown in by (g) and (h). 84

Figure 5.5 The performance of degradation classification for four ancient document images by two metrics, VDQAM and MDQM. The first raw (e(i)-e(iv)) and second raw (f(i)-f(iv)) bars show the MDQM and VDQAM operation for the probability estimation of different distortions types, respectively. The X-axis of bars demonstrates different distortion types in the degraded images of MHDID dataset: paper transluency (PT), readers' annotations (RA), stains (S) and worn holes (WH). 88

Figure 5.6 The histogram of run-time (seconds/image) and accuracy for seven NR-IQA metrics. 93

Figure 6.1 Three DMD images with partial segmentation. Text, non-text and graphical regions are shown with blue, yellow and red colors, respectively. 99

Figure 6.2 An original DMD image (*a*), its grayscale edge magnitude with Sobel operator (*b*) and color saliency based edge map (*c*). 102

Figure 6.3 An original DMD image (a), its standard deviation image (b) and standard deviation of normalized image (c). 103

Figure 6.4 Color segmentation results 103

Figure 6.5 Results of segmentation into three layers of graphics, text and non-text. 104

Figure 6.6 Histograms of MSCN coefficients for three layers of the two DMD images with different MOS values. (c) MSCN distribution of three layers in Img #1; (d) MSCN distribution of three layers in Img #2. 105

Figure 6.7 Comparing MSCN of gradient histograms vs MSCN of color gradient saliency histograms for three DMD images. 106

Figure 6.8 The chart of feature extraction time (in seconds) for seven IQA methods. 110

LIST OF ABREVIATIONS

AGGD	Asymmetric General Gaussian Distribution
BQAD	Blind Quality Assessment for DMDs
CC	Correlation Coefficient
CG	Color Gradient
CF	Colorfulness
DMD	Degraded Medieval Document
DMDD	Degraded Medieval Document Dataset
DNT	Divisive Normalization Transformation
DS	Distortion Specific
ÉTS	École de Technologie Supérieure
FR-IQA	Full-Reference Image Quality Assessment
GS	Gradient Similarity
GGD	General Gaussian Distribution
HDI	Human Document image
HVS	Human Visual System
IQA	Image Quality Assessment
KRC/KRCC	Kendall Rank Correlation Coefficient
LWMPA	Locally Weighted Mean Phase Angle
MATLAB	Matrix laboratory

MSCN	Mean Subtracted Contrast Normalized
MHDID	Multi-distortion Historical Document Image Database
MDQM	Multi-distortion Document Quality Measure
MOS	Mean Opinion Score
NR-IQA	No-Reference Image Quality Assessment
NDS	Non Distortion Specific
NSERC	Natural Sciences and Engineering Research Council of Canada
OCR	Optical Character Recognition
PCR	Pair Comparison Rating
PCC/PLCC	Pearson Linear Correlation Coefficient
RR-IQA	Reduced-Reference Image Quality Assessment
STD	Standard Deviation
SVR	Support Vector Regression
SVM	Support Vector Machine
SI	Spatial Information
SRC/SRCC	Spearman Rank-order Correlation Coefficient
VN	Variance Normalized
VDIQA	Visual Document Image Quality Assessment
VDQAM	Visual Document image Quality Assessment Metric

INTRODUCTION

Heritage is our legacy from the past, what we live with today, and what we pass on to future generations (United Nations Educational, Scientific and Cultural Organization (UNESCO)). Damaged manuscripts and documents constitute an important part of the heritage. They are the memory of human cultures, their history, their achievements, their lifestyle, their individual and social behaviors (Hedjam *et al.*, 2015). Recent years have seen increasing efforts in archiving and digitizing these valuable documents in order to safeguard them against deterioration, and to process them by signal processing techniques (Antonacopoulos & Downton, 2007; Hedjam *et al.*, 2015; Savino & Tonazzini, 2016). Having a large amount of the documents archived, it is very interesting to use automatic techniques to extract and use meaningful information from their images. Because of aging, improper handing and environmental factors, these documents suffer from low to high degrees of degradation. These physical degradations appear on the digitized document images as well. According to the taxonomy of (Lins, 2009), these types of degradation refer to the “physical noises" in document images. The following, is a list of the physical noises (Lins, 2009): folding marks, paper translucency, paper aging, paper texture, paper punching, stains, torn-off regions, worm holes, readers’ annotations, physical blur, carbon copy effect, scratches and cracks, sunburn, and inadequate printing. Physical noises can greatly affect the performance and accuracy of the document processing algorithms.

In order to reduce the effect of the physical noises on the document processing algorithms, it is common to automatically remove these noises. The performance of the current methods to remove physical distortions is far beyond the HVS ability to distinguish type and severity of degradations. Given the strong ability of the HVS to identify degradations in document images, it is very interesting to develop degradation assessment metrics that mimic the HVS. In turn, enhancement methods or other processing algorithms can be developed or tuned with respect to these metrics. Visual document image quality assessment, if not used by the enhancement methods, can have other applications. It helps in the classification of the document images on the basis of their visual quality. For example, an extremely degraded document can be picked and treated manually in order to extract maximum information or to prevent further document deterioration. The performance expectation of the document processing algorithms can be estimated through visual document image quality assessment (VDIQA) metrics.

The applications of VDIQA and the hypothesis that a document image can be assessed even though it cannot be enhanced are good motivations to develop VDIQA datasets and metrics. There are few document image quality assessment (DIQA) datasets available in the literature, and most of them are evaluated based on the OCR performance. However, OCR engines are not perfect, specially for some of the languages, old writing styles and fonts. Thorough this thesis, we use the term VDIQA when human judgments are used and DIQA when OCR accuracy is used.

The quality of images can be assessed subjectively and objectively. Subjective evaluation is more accurate and ecologically valid. However, it involves human participation; therefore it is time-consuming and expensive. To overcome these limitations, objective metrics have been proposed in the literature. Objective IQAs are mathematical models that approximate the results of subjective IQA. The main goal of objective IQA is to supply quality metrics that can automatically predict perceived image quality. According to the availability of non-distorted images, objective image quality assessment can be classified into three categories: full reference (FR)(Chandler & Hemami, 2007; Wang *et al.*, 2004; Zhang *et al.*, 2011; Sheikh & Bovik, 2006; Nafchi *et al.*, 2016), reduced-reference (RR)(Wang *et al.*, 2006; Rehman & Wang, 2012; Li & Wang, 2009), and no-reference (NR) IQA models (Moorthy & Bovik, 2011; Saad *et al.*, 2012; Mittal *et al.*, 2012; Liu *et al.*, 2014; Gu *et al.*, 2016). In NR-IQA, only distorted images are available. This thesis focuses on historical document images, for which the reference image is not available.

NR-IQA metrics perform according to the statistical regularities of natural images in spatial and transformed domains. The deviation between statistical regularities of distortion-free and distorted images is considered in the design of the NR-IQA models. None of the available metrics are designed for quality assessment of the physically distorted document images. Basically, they do not access the distortions appearing on the documents because of aging and physical condition. Also, majority of these metrics work on text regions to correlate with OCR accuracy. This opens up the main question of this thesis, **what are effective strategies and image features to assess and classify physically degraded documents that provide a fair correlation with the human perception of document image quality?**

To answer this question, it is necessary to study the main challenges toward assessment and classification of the physical distortions in document images. These challenges are detailed as follows.

0.1 Problem Statement (PS)

General image quality assessment metrics have improved a lot in recent years, but none of them are designed to deal with the physical distortions of old document images. There are many types and severity of physical distortions. Physical distortions may have irregular size and usually appear locally on the documents. In the following, major challenges and problems toward assessment of physical distortions are explained.

0.1.1 PS1. No dataset

To the best of our knowledge, there is no dataset of old and degraded document images consisting of the full-page color images with associated MOS values in the literature. Therefore, it is difficult to propose or validate NR-DIQA models, especially if those models need to be trained.

0.1.2 PS2. Physical degradations

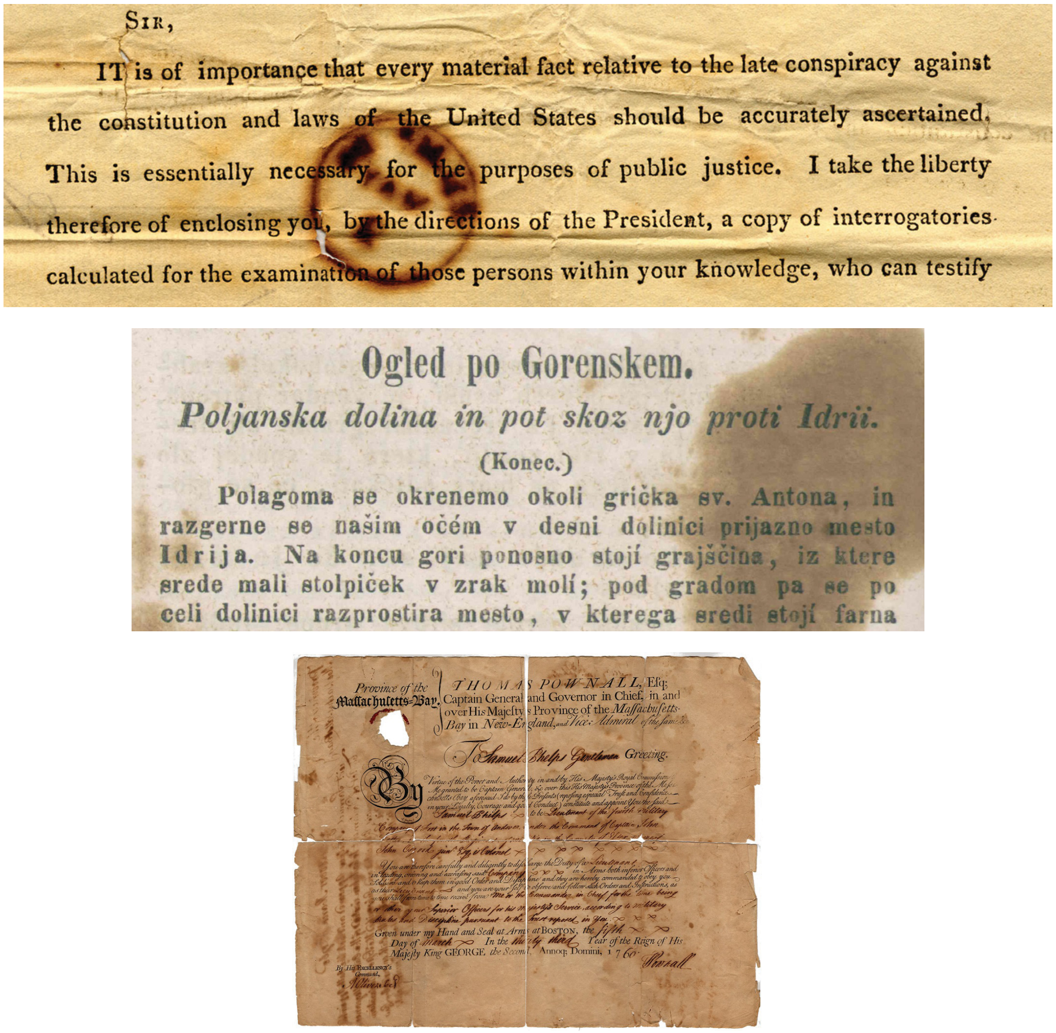
Ancient documents suffer from different types of degradations such as bleed through, show through, alien ink, stain, paper deteriorations, etc. In Figure 0.1, a few examples are shown. There are several degradation types in historical documents that we cannot often see in modern documents. Bleed through and stain are just two examples, among others (Moghaddam & Cheriet, 2010). For better understanding of their difference, Table 0.1 lists a few degradation types in modern and historical document images. Both modern and ancient documents may have any types of degradations listed here, but ancient documents are more likely to have physical distortions. Therefore, source and type of noises are often quite different for modern and ancient documents. Assessing physical distortions is not easy because of their irregular size and pattern. Also, given the local nature of the physical distortions, the impact of these distortions on the whole document quality is difficult to be modeled.

Table 0.1 Selected degradation types for modern and ancient documents.

Document	**Type of degradation**
Modern	Blur, contrast, white noise, compression effects, denoising effects, transmission errors, salt and pepper noise.
Ancient	Bleed/show through, paper deteriorations, low resolution, illumination variation, weak strokes, different ink distortions, interfering patterns.

0.1.3 PS3. Limited research

In the design of DIQA metrics for modern documents, it is common to directly borrow features and knowledge from the literature of IQA. However, these features cannot be *directly* used for IQA of physically distorted documents. Statistical characteristics of ancient document images with physical distortions are not studied in the literature.



SIR,

IT is of importance that every material fact relative to the late conspiracy against the constitution and laws of the United States should be accurately ascertained. This is essentially necessary for the purposes of public justice. I take the liberty therefore of enclosing you, by the directions of the President, a copy of interrogatories calculated for the examination of those persons within your knowledge, who can testify

Ogled po Gorenskem.

Poljanska dolina in pot skoz njo proti Idrii.

(Konec.)

Polagoma se okrenemo okoli grička sv. Antona, in razgerne se našim océm v desni dolinici prijazno mesto Idrija. Na koncu gori ponosno stojí grajščina, iz ktere srede mali stolpiček v zrak molí; pod gradom pa se po celi dolinici razprostira mesto, v kterega sredi stojí farna

Province of the
Maſſachuſetts-Bay.

THOMAS POWNALL, Eſq;
Captain General and Governor in Chief, in and over His Majeſty's Province of the Maſſachuſetts-Bay in New-England, and Vice-Admiral of the ſame &c.

To Samuel Phelps Gentleman Greeting.

By Virtue of the Power and Authority in and by His Majeſty's Royal Commiſſion to Me granted to be Captain General, &c over this His Majeſty's Province of the Maſſachuſetts Bay aforeſaid I do by theſe Preſents (repoſing eſpecial Truſt and Confidence in your Loyalty, Courage and good Conduct) conſtitute and appoint You the ſaid Samuel Phelps to be Lieutenant of the fourth Military Company of Foot in the Town of Andover, under the Command of Captain John [illegible] in the County of Eſsex whereof John Osgood jun.r Eſq. is Colonel

You are therefore carefully and diligently to diſcharge the Duty of a Lieutenant in leading, ordering and exerciſing said Company in Arms both inferior Officers and Soldiers and to keep them in good Order and Diſcipline and they are hereby commanded to obey you as their Lieutenant and you are your Self to obſerve and follow ſuch Orders and Inſtructions, as you ſhall from time to time receive from Me or the Commander in Chief for the time being or other your Superior Officers for his Majesty's Service, according to military Rules and Discipline pursuant to the Trust reposed in You.

Given under my Hand and Seal at Arms at BOSTON, the fifth Day of March In the thirty third Year of the Reign of His Majeſty King GEORGE the Second, Annoq; Domini, 1760.

Pownall

By His EXCELLENCY's Command,
A Oliver Secr

Figure 0.1 Examples of ancient document images with different types of degradation.

0.1.4 PS4. No quality assessment metric for historical documents

To the best of our knowledge, no metric is proposed for quality assessment of physically degraded documents in the state of the art, while many blind metrics are proposed for quality evaluation of natural images in the literature. It is predictable that these available quality assessment metrics are not functional for assessing historical document images.

0.1.5 PS5. No degradation identification model for HDIs

In the literature, there is not any degradation classification model in order to detect and estimate the distortion type and severity of the physically distorted documents. This is connected to the lack of a dataset with labeled degradations. Such models provide a priori information for many document processing applications.

0.2 Research questions

In order to address the aforementioned problems and drive our work methodology, we further detail the problem statement into three research questions (RQ) as follows. The answer to these three research questions will be provided with details in our papers (Chapter 3-6).

0.2.1 Research Question (RQ1)

1. How to develop a dataset with associated MOS values for HDIs?
2. How to extract statistical features from HDIs?
3. How to segment an ancient document into several meaningful layers?
4. What is a good strategy to design a blind quality metric for HDIs?

0.2.2 Research Question (RQ2)

1. How to develop a dataset with different category of physical distortions for ancient documents?
2. What statistical features are more efficient for designing a DIQM?
3. How to design a more effective yet more efficient quality metric for HDIs?
4. Are extracted features effective at classifying degradation types?

0.2.3 Research Question (RQ3)

1. How to develop a dataset for DMDs?
2. How to segment a DMD image into several meaningful layers?
3. How to design a quality assessment metric for DMD images?

0.3 Contributions

There are limited researches on document image quality assessment which none of them have taken into account the physical distortions of degraded document images. This means that there is no dataset, quality assessment metric and degradation classification model of physically distorted document images in the literature. **Therefore, the purpose of this thesis is to develop HVS-based datasets, propose efficient quality assessment metrics for ancient degraded documents and design a degradation classification model for the physically distorted document images.** The majority of the thesis' contributions belong to the field of no-reference document image quality assessment where there is no pristine-quality reference document image available.

Firstly, there are two contributions in our first research work that are explained as follows. First, VDIQA dataset is created based on HVS judgments. There are a few DIQA datasets available in the literature which are based on the OCR performance. Due to the limitations of OCR performance on different languages and etc, human judgments are used in the developed dataset instead of the OCR accuracy. Second, an image quality assessment metric is proposed to provide a quality score for the historical document images with high correlation with MOS values. The proposed metric works on four different layers of each degraded document. Indeed, statistical features are extracted from each segmented layer. Segmentation into four layers is done by the analysis of Log-Gabor filter. The ability of several no-reference IQA metrics is evaluated on the developed database (VDIQA). Experiment results show that our method has considerable performance improvements and a moderate run-time efficiency in comparison with other NR-IQA metrics.

Secondly, we focus on the creation of a new dataset for degraded document images because the number of images in the VDIQA dataset is not sufficient for some learning-based tasks such as degradation classification and etc. Moreover, no information about the types and probability of distortions is provided for this dataset. Therefore, MHDID is introduced with four categories of distortions and more number of degraded images in comparison with VDIQA dataset. This dataset is useful for the purpose of degradation classification and modeling. It should be mentioned that the PCR method is utilized as a subjective rating method for evaluating the visual quality of degraded document images.

Thirdly, measuring the amount of degradation and quality assessment of degraded documents is highly desirable for applications such as selecting the proper algorithms for enhancement and analysis of document images, filtering the damaged images, tuning the processing algorithms parameters, document repairing, psychological study, etc. The first contribution of this work is the proposition of MDQM metric for quality assessment of physically degraded document images. This metric is based on three sets of spatial and frequency image features. These features are extracted from two layers of text and non-text and mapped to the MOS values using regression function. The second contribution of this work is to estimate the probability of four common distortion types in the degraded images. In our experiment, the correlations of seven NR-IQA metrics with the MOS values are evaluated on two available datasets (VDIQA and MHDID). It is shown that the performance of MDQM metric is significantly better than the state-of-the-art NR-IQA metrics. Moreover, the experimental results demonstrate that MDQM metric does not only lead to a high efficacy for classification of the various degradations but also maintains a remarkable run-time efficiency.

Lastly, a new dataset (DMDD) and a quality assessment metric (BQAD) are proposed for degraded medieval document images. The same as the creation of the VDIQA and MHDID datasets, the PCR method is used for pair comparison of degraded medieval documents in the DMDD. In the developed dataset, each DMD image contains three parts: text, non-text and graphic area. A color saliency approach a phase-based binarization method are used for the segmentation of each DMD image into three mentioned parts. Then, statistical features are extracted from each segmented layer and map to MOS values by a regression function. The experimental results performed on the DMDD dataset demonstrate that the proposed blind quality metric has a good correlation with the HVS and also its performance is quite better than the state-of-the-art blind metrics. It should be mentioned that DMDD is the first dataset published for quality evaluation of DMD images. Also, the BQAD metric is the first attempt to evaluate the quality of DMD images.

Table 0.2 lists every contribution of the three mentioned research works in this thesis.

Table 0.2 All contributions in the three research works

	Contributions	**Type of document**
First research work	VDIQA dataset and VDQAM metric	Historical document images
Second research work	MHDID dataset, MDQM metric and a degradation classification model	Historical document images
Third research work	DMD dataset and BQAD metric	Degraded medieval documents

0.4 Structure of the thesis

In this thesis, we focus on dataset's creation, image quality assessment and degradation classification of ancient document images, its challenges, and solutions that we bring to tackle these challenges. This work is structured as follows:

- In **Literature review**, we discuss the existing state-of-the-art datasets, subjective and objective image quality assessment and degradation classification models for different types of images. More details about the state of the art are provided in chapters 3 to 6 concerning the article publications.

- In **General Methodology**, we consider the problems, their solutions, and used techniques in our work. Indeed, a comprehensive description of the proposed methods is provided for readers that can understand how we tackle the objectives posed in this work.

- **Article publications** are four chapters dedicated to our journal and conference publications. In these chapters, two datasets of physically degraded old documents are developed in this work in order to train and validate DIQA metrics. Also, two proposed no-reference metrics for quality assessment of historical document images are described. In addition, a method is proposed to classify four types of physical degradations. Finally, a new dataset and a quality assessment metric are explained for degraded medieval document images.

- Chapter **General Discussion** provides a general discussion on the drawbacks and weaknesses of the proposed methods.

- Finally, **General Conclusion and Future Works** summarizes the work accomplished in this thesis and proposes some suggestions for future works.

CHAPTER 1

LITERATURE REVIEW

This chapter presents a review of the state-of-the-art methods related to the subjective, objective quality assessment metrics and degradation identification algorithms. This chapter is divided into three sections that are in line with the challenges discussed in the problem statement. The first section explains about image quality assessment with details. The second section discusses on different degradation identification algorithms for natural and ancient images. The third section covers several evaluation measures for objective quality metrics.

1.1 Image quality assessment

Quality of images can be assessed subjectively and objectively. The most accurate way to judge the quality of images and videos is to conduct subjective experiments. However, given the considerable amount of visual data, such experiments are very time-consuming. Indeed, subjective evaluation is more accurate, ecologically valid but involves human participation. Therefore, automatic objective image/video quality assessment metrics that can mimic the subjective evaluation of visual data are of great interest. These computational models take into account changes in visual data information only if these changes cause annoyance for viewers. The non-visible information changes in visual data are ignored by these metrics. In the following, more details about the subjective and objective image quality assessment will be provided.

1.1.1 Subjective image quality assessment

Subjective image quality assessment methods could be classified into two categories: direct scaling methods and indirect scaling methods (Ferwerda, 2008). In comparing these two methods, indirect scaling methods provide higher discriminatory power and can be less complicated and tiring for the subjects. Also, it should be mentioned that indirect methods need the lower number of observers to provide the same reliability in comparison with direct scaling methods. In the following, more explanations about these two scaling methods will be provided.

1.1.1.1 Direct scaling methods

Direct scaling methods is based on the judgments of subjects which obtain from each particular stimulus directly. When the results from all subjects are collected, the outlier results are detected and deleted. The final outcome of the experiments is named MOS values. Three direct scaling methods will be explained in the following.

- Absolute category rating (ACR): is considered as the simplest subjective method. In this method, subjects are asked to select one of the five grade scales. These scales are sorted by quality in decreasing order: 5.excellent, 4.good, 3.fair, 2.poor and 1.bad. The stimuli are shown to the subjects in a random order one at a time. MOS values with ranging from 1 to 5 are calculated from the results of this experiment. In order to increase the discriminatory power and reliability, stimuli can be displayed to subjects repeatedly, because ACR method is the least time consuming.
- Degradation category rating (DCR): in this method, we assume that the stimulus is degraded compared to the reference. The main purpose of this subjective method is to evaluate how much does the distortion affect the perceived quality. The subjects are asked to select one of the five levels of degradation in each degraded image: imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying. It worth to mentioned that the length of the test is more in comparison with ACR method. The same as ACR method, MOS values with ranging from 1 to 5 are calculated from the results of this experiment.
- Double stimulus continuous quality scale (DSCQC): the last direct scaling method mentioned in this thesis is DSCQC. The stimuli are displayed to observers in pairs. However, the observers are not explicitly told that one of the two stimuli is reference and they are supposed to evaluate both of them at the same time. The recommendation allows two different variants (Krasula, 2017):

  Variant I: The observer is allowed to freely switch between the two stimuli and then score the quality of both of them on the scale.

Variant II: The stimuli are presented twice.

1.1.1.2 Indirect scaling methods

As mentioned before, indirect scaling methods are the sequential-task, more comfortable, and less demanding for the subjects in comparison with direct scaling methods. In this thesis, two most popular approaches are described namely ranking and paired comparison rating (PCR).

- Ranking: the process of ranking is self-explanatory. A set of degraded documents, which are obtained from the same source content (with or without the presence of the reference), are displayed to subjects. Then, observers are asked to rank them. The procedure of ranking method is very popular especially in printing industry where it is easy to provide the observers with access to all of the pair images in the set at the same time (Krasula, 2017). In (Kumar & Ramakrishnan, 2011), a dataset is introduced based on the ranking method. In this paper, five scales are described as: 1. Image with degradations caused by scanning; 2. Highly degraded document image; 3. Background degraded document image; 4. Slightly degraded document image and 5. Good document image.

- Pair comparison rating (PCR): it has been introduced to overcome some limitations of the standard subjective IQA approaches, for example, the fact that an arbitrary score scale has to be defined. This may lead to ambiguous scores in many cases. PCR also offers the possibility to consider stimuli with a similar level of quality. At each trial, two document images are shown to the subjects. The subjects are asked to compare the two images. A score of 1 is assigned to the image with higher quality and -1 to the other. If the observer could not perceive any difference between the two images, a score of 0 is assigned to both images. In (Obafemi-Ajayi & Agam, 2012), the subjects are asked to make a decision about the quality of each pair images: left or right image is better or if they seem to be of identical quality.

Another very important way of creating datasets of document images is to use OCR accuracy. In the following, available datasets for IQA and DIQA are provided.

1.1.2 Subjective document image quality assessment and datasets

There are few DIQA datasets available in the literature, and most of them are evaluated based on the OCR performance. To the best of our knowledge, there is no existing DIQA datasets based on the human perception in state-of-the- art. The majority of standard IQA datasets is available for natural images. It should be mentioned that two types of indirect scaling approaches, which were explained with details in previous sections, have been utilized to develop a subjective DIQA: ranking and pair wise (Obafemi-Ajayi & Agam, 2012; Kumar & Ramakrishnan, 2011).

IQA datasets of natural images (Sheikh *et al.*, 2006; Ponomarenko *et al.*, 2013), synthetic images (Kundu & Evans, 2015), photo retouched images (Vu *et al.*, 2012), and screen content images (Yang *et al.*, 2015) are publicly available in a fast growing field of research. However, there has been little effort to develop datasets and metrics for quality assessment of the document images. The majority of the datasets that have been introduced are either not available or not available to the public (Chou & Yu, 1993; Govindaraju & Srihari, 1995; Kulesh *et al.*, 2001; Obafemi-Ajayi & Agam, 2012). The very few quality metrics for DIQA are either not available or not available to the public (Chou & Yu, 1993; Cannon *et al.*, 1999; Park *et al.*, 2000; Kumar & Ramakrishnan, 2011; Obafemi-Ajayi & Agam, 2012; Kulesh *et al.*, 2001). The authors of (Eilertsen *et al.*, 2013) surveyed DIQA/VDIQA metrics and datasets. The DIQA datasets can have different characteristics. Images in a dataset might be in the form of color, gray-level, or binary. Each image may show a character, a word, a sentence, or a full page.

OCR accuracy and MOS are two data types available in these datasets for the evaluation purposes. The majority of the VDIQA datasets are in the form of binary images (Chou & Yu, 1993; Govindaraju & Srihari, 1995; Obafemi-Ajayi & Agam, 2012). The dataset introduced in (Kumar *et al.*, 2012) consists of 135 gray-level images with blur distortion. The MOS values for this dataset are collected and computed by crowd-sourcing. This dataset consists of camera-captured document image with varying levels of focal blur introduced manually during capture. Three different OCR engines (ABBYY Finereader, Tesseract and Omnipage) are used to obtain character level OCR accuracy for each image. A dataset for assessing the quality of scanned document images is proposed by (Blando *et al.*, 1995). In this dataset, two sets of test data are used for the experiments and six OCR systems are utilized to process their data sets and collected character accuracy for each image.

1.1.3 Objective image quality assessment

The main goal of objective IQA is to supply quality metrics that can automatically predict perceived image quality. These metrics mimic the quality predictions of human observers since the human eyes are the ultimate viewer. IQA models can be used in parallel with an image processing system to provide feedback for the system or can be directly embedded into the system. Performance of a recognition system for an application can be greatly affected by image distortions. Objective IQA can help to estimate the performance expectation of a recognition system or can provide information to preprocess the input image first and run the recognition system on the processed image.

Objective image quality assessment (IQA) models can be categorized into full-reference (FR), reduced-reference (RR), and blind/no-reference (NR) depending on their access to the reference image with pristine quality. Figure 1.1 provides an illustration of this categorization. Both reference image and possibly distorted image are available for an FR-IQA metric. RR-IQA models have full access to the distorted image and perform assessments with respect to some certain statistical properties of the reference image. To perform a quality assessment of a possibly distorted image, NR-IQA models have no access to the reference image.

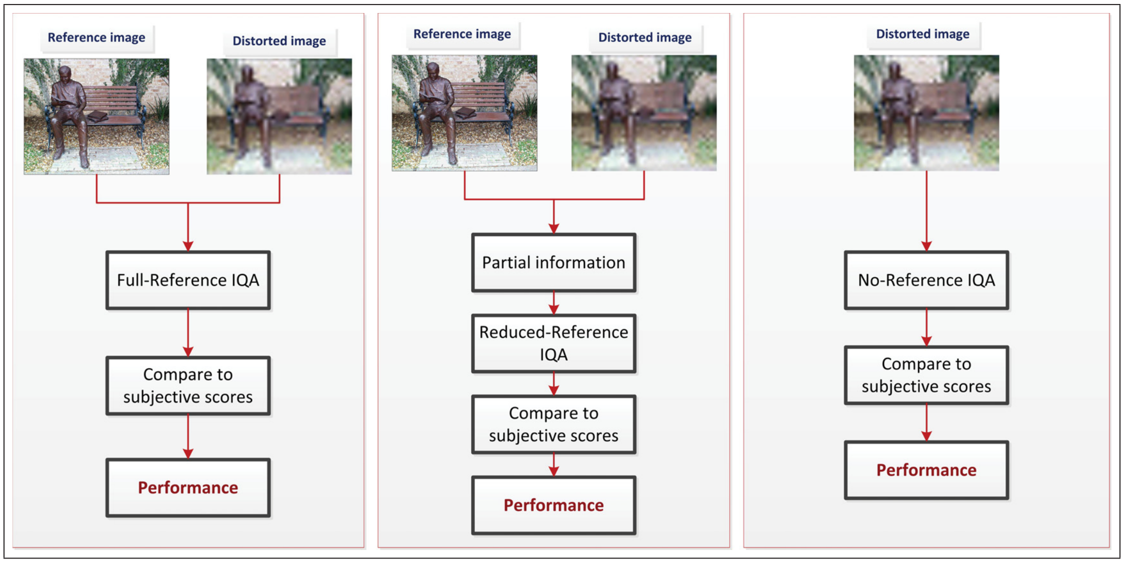


Figure 1.1 Three different objective image quality assessment models. From left to right, full-reference IQA model, reduced-reference IQA model, and no-reference IQA model.

1.1.3.1 FR (Full Reference)

Full reference requires the reference image in order to evaluate the quality of the degraded images. The design of FR-IQAs is mainly based on measuring the difference between extracted features from the reference and distorted images. It seems that efforts in the literature on FR-IQAs are focusing on designing more robust metrics to different image distortions and metrics with higher efficiency at speed and memory. The flowchart of typical FR-IQA models is shown in Figure 1.2.

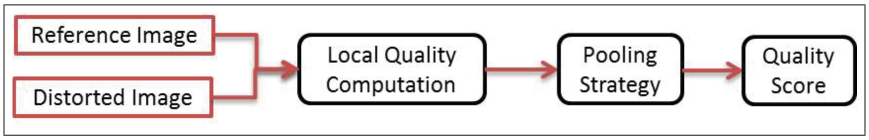


Figure 1.2 The flowchart of FR-IQA models.

1.1.3.2 RR (Reduced Reference)

In reduced reference (RR) image quality assessment, partial information of the reference image is essential to compute the visual quality of distorted images. Appropriate RR features should provide an efficient summary of the reference image and also should be sensitive to a variety of image distortions. In the literature, there are some methods for RR image quality assessment that some of them will be illustrated as the following.

- (Wang & Simoncelli, 2005) proposed a RR image quality assessment method based on a natural image statistic model in the wavelet domain. Kullback Leibler was used in order to measure the distance between the marginal probability distributions of wavelet coefficients of the reference and distorted images. Then, the marginal distribution of the coefficients in individual wavelet subbands was fitted by GGD in order to summarize the marginal distribution of wavelet coefficients of the reference image. Therefore, because the measurement is based on marginal distributions of wavelet coefficients, the method is insensitive to small geometric distortions such as spatial translation, rotation and scaling. This method is computationally efficient and utilizes a few parameters. Also, this method works very well on several types of distortions, in spite of application-specific methods that work for some distortion types.

- (Li & Wang, 2009)proposed another RR metric that is based on a divisive normalization image representation. Divisive normalization transformation (DNT) is computed by a Gaussian scale mixture statistical model of image wavelet coefficients. Then, the quality of distorted images is evaluated by comparing a set of RR features extracted from DNT-domain representations of the reference and distorted images. Due to the fact that DNT image representation has simultaneous perceptual and statistical relevance and its statistical properties are significantly changed under different types of image distortions. Therefore, these properties make it well-suited for the development of RR-IQA algorithms.

1.1.3.3 NR (No-Reference)

In many practical computer vision tasks, no perfect version of the distorted images exist, thus these tasks require no-reference image quality assessment (NR-IQA). NR-IQA (Wang *et al.*, 2002; Zhou *et al.*, 2008; Mittal *et al.*, 2012; Moorthy & Bovik, 2011) is difficult task, because NR-IQA models evaluate the quality of the images without accessing to original images. NR metrics are highly desirable in ancient documents (maybe not in their current forms), because original images are not available.

NR-IQA based on the prior knowledge of the distortion type, can be categorized into distortion-specific (DS) and non-distortion-specific (NDS) (Zhang & Chandler, 2013). The non-distortion specific NR-IQAs usually follow a training/learning based approach or natural scene statistics (NSS). In DS, the type of distortion is known (such as JPEG, JPEG2000, white noise, etc). In NDS, there is no information about the type of distortion. NR-IQAs follow two strategies. In the first strategy, image features are extracted and relation between these features and MOS is modeled by a function. This function is then used to map features that are extracted from unseen distorted images to MOS. In the second approach, natural statistics of images are modeled. The metrics in this category measure the amount of difference between the statistics of some images with ideal quality, and those of the distorted images. The more difference means more severe distortion. It is common in NR-IQAs that distortion type is first determined and quality score is computed accordingly.

In general, mathematical models utilized by NR-IQAs are more complicated than those for FR-IQAs. The features used by FR- and NR-IQAs are somewhat similar. Gradient, Laplacian, and sparse features are widely used by both, while the most common pooling strategy for RR and NR-IQAs is the percentile pooling. Mean pooling, weighted mean pooling, and recently deviation pooling are common pooling strategies used by FR-IQAs. In terms of efficiency, NR-IQAs have the advantage of only processing the distorted images. However, NR-IQAs are in general much slower than the FR-IQAs. The reason is that NR-IQAs are trained with many extracted features and probably in different scales (might be the case for FR-IQAs as well). Also, it is common that some preprocessing steps are used by NR-IQAs on the input images to remove dependency of features which are time-consuming. Most successful approaches toward this challenge utilize the Natural Scene Statistics (NSS) based features. Some of the popular NSS based NR metrics such as BRISQUE (Mittal *et al.*, 2012) and DESIQUE (Zhang & Chandler, 2013) will be explained here.

- BRISQUE (Blind/Referenceless Image Spatial QUality Evaluator), (Mittal *et al.*, 2012) proposed a simple, efficient and effective method based distortion generic blind/no reference (NR) quality assessment, which works in the spatial domain. The main goal of this method is to build a computational model to automatically predict human perceived image quality without a reference image and without knowing the distortion present in the image. The distribution of MSCN is always Gaussian. Since MSCN coefficients are definitely homogenous, therefore the signs of these coefficients exhibit a regular structure. BRISQUE models this structure using the empirical distributions of pairwise products of neighboring MSCN coefficients in four directions: horizontal, vertical, main diagonal and secondary diagonal. After extracting 36 features (18 at each scale), BRISQUE employs two stage framework for training: distortion identification and distortion specific quality assessment. In this approach, the same set of features are used to identify the distortion afflicting the image as are used for distortion specific quality assessment. BRISQUE is computationally efficient and provides acceptable quality predictions for several types of distortions.

- DESIQUE (DErivative Statistics-based Image QAuality Evaluator), (Zhang & Chandler, 2013) extracted statistical features in two domains: spatial and frequency, because the perceptual quality can be influenced by both the spatial and frequency information in an image.

  In the spatial domain, luminance values of an image are modeled into point wise based statistics and pairwise based log-derivative statistics. i) *Point-wise based statistics* is considered for the relationship of pixel pairs. MSCN coefficients was modeled by a zero mean general gaussian distribution (GGD). ii) *Pairwise based log derivative statistics* are formed for the relationship of pixel pairs based on log-derivative statistics. Five types of log-derivatives features are utilized in order to model the relationship between pixel pairs of MSCN coefficients.

In order to estimate the quality based on frequency domain, image is decomposed into horizontal and vertical orientations by using log-Gabor filter. Then, five types of log-derivatives features are extracted from each orientation.

After extraction some features in two domains, DESIQUE similar to BRISQUE (Mittal *et al.*, 2012), uses a two-stage framework: distortion identification and distortion specific quality assessment. In distortion identification, a support vector regression (SVR) is trained in order to calculate the quality of the distorted images. SVR also used in distortion specific quality prediction in order to map the feature vectors to an associated quality score. After extraction of aforementioned features in two domains, a classification model is trained by a support vector classification machine to measure the probability that each distortion type exist in certain distorted images. Then, for each distortion type a particular regression model is trained by SVR in order to map certain feature vector to the associated quality score. The performance of DESIQUE is shown to be better than BRISQUE.

1.1.4 Objective document image quality assessment

The goal of objective DIQA is to develop a computational model that can predict the quality of a document image automatically and accurately (Eilertsen *et al.*, 2013). Objective document image quality assessment models that are designated solely for the assessment of document images will provide much better image adjustments. That means special purpose DIQAs can lead to the proper adjustment of document images that are in higher correlation with the human visual system.

If final consumer of document image is a machine, OCR accuracy usually evaluate performance image quality assessment. However, if human eyes are the final consumer of document images, it is human perception that judges on the quality of images (see figure 1.3).

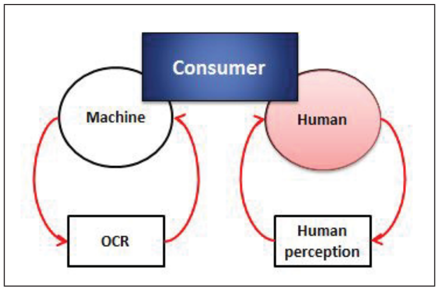


Figure 1.3 Overview on document IQA models

1.1.4.1 OCR accuracy

Objective DIQA based on the OCR accuracy is categorized into content based and degradation based. In content based, the main focus is on inherent property of the document, while a degradation based approach focuses on evaluating the different degradations that arise from the production process, which are independent of the content of the document (Ye & Doermann, 2012).

- Assessment based on the content: In (Chou & Yu, 1993), a quality metric is proposed to measure the quality of handwritten Chinese characters. In this work, if the quality of a character is similar to the average of a large volume of that character category, its quality is considered “good". Handwritten characters from a specific category are sorted by their distances to the average sample of that category. In order to sort their quality, three types of features are extracted: stroke-density distribution of a Chinese character, Histogram and Crossing Count.
- Assessment based on the degradation: In (Ye & Doermann, 2012), an unsupervised feature learning metric was proposed to assess the quality of degraded document images. The main purpose of this work is to propose a computational metric to predict OCR accuracy of a gray-scale document image automatically. In this method, raw-image-patches are extracts from a set of unlabeled images to learn a dictionary in an unsupervised manner. Given an image, a set of raw-image patches are extracted as local features. Then, the dictionary using soft-assignment encoding with max pooling is used

for encoding to obtain effective image representations for quality estimation. Finally, SVR is used in order to map the image features to an image quality score (Eilertsen *et al.*, 2013). In (Peng *et al.*, 2011), an automated image quality assessment metric is introduced to predicts the degradation degree of the camera-captured document images according to their statistic features. It is worth to mention that this proposed metric quantifies several degradations and accurately predicts the impact on OCR error rate.

1.1.4.2 Human perception

Proposing a quality metric based on the human perception is a momentous task in abundant applications. A character-based automated human perception quality assessment is proposed for document images in (Obafemi-Ajayi & Agam, 2012). In this work, three types of features (morphological-based features, noise removal-based features and spatial characteristics) are computed from character images to obtain a measure of degradation quality for character images. Also, a two-stage systems is proposed to estimate human perception based on the level of distortions. In the first stage, this system contains a degradation classifier that distinguishes the type of degradation. The second stage consists of a set of two regressors that appropriate predictor based on the type of distortion is selected. A full-reference DIQA method is proposed by (Lu *et al.*, 2004). This method is a distance-reciprocal distortion (DRD) measure for binary document images. This method is based on the observation that the distance between two pixels plays an important role in their mutual interference perceived by subjects (Eilertsen *et al.*, 2013). In this paper, all the test images are divided into four groups based on the DRD values. In each trial, four test images which are chosen from four groups, are shown to observers and they are asked to rank the visual quality of the four images. It should be mentioned that the smaller ranking score indicates less distortion. The ranking scores collected from the 60 subjects are analyzed and compared with the rankings based on the average DRD scores. Finally, 240 scores are obtained by the 60 observers for four groups of document images.

1.1.5 Discussion

The majority of IQA datasets in the literature are designed for natural images. Also, there are some DIQA datasets based on the OCR accuracy in state of the art which some of them are not available. As mentioned in contribution's section, OCR accuracy is not enough reliable in comparison with human perception to create a dataset. Also, most of the existing DIQA metrics and datasets were proposed for a special type of degradation in the literature.

Generally, the conclusion derived from this part of the literature review is that an objective quality evaluation metric for degraded document images is not considered in state-of-the-art. Furthermore, no subjective quality assessment for historical document images consisting of the full-page color degraded images with associated MOS values has never been taken into account in the literature.

1.2 Distortion identification algorithm

Degradation classification and characterization is a comparatively new area of research. The researchers in our field agree that there is no universal algorithm that can efficiently estimate all of the existing degradations in images. In this section, we briefly review the degradation classification models which are designed for the limited degradations in natural and historical document images in the literature.

1.2.1 Natural images

In order to propose an objective IQA metric, an algorithm is designed to estimate the presence of five common noises in natural images (Moorthy & Bovik, 2010). These degradations are those from LIVE database (Sheikh *et al.*, 2006): JPEG, JPEG2000 (JP2K), white noise (WN), Gaussian Blur (Blur) and Fast fading (FF). Two stages constitute the structure of this algorithm: a classification and evaluation models. Let $p_i, i = 1, ..., 5$ and $q_i, i = 1, ..., 5$ demonstrate the probability of five noises in an image and the quality scores from each of the five quality assessment algorithms (corresponding to the five distortions), respectively. Finally, the quality of an image is obtained with a probability-weighted summation. It should be mentioned that a multi-class SVM classifier is utilized to classify each image into one of five degradation categories. It is likely that some of these classes overlap to some extent. In this case, a greater value shows a higher proportion of that noise in the image.Indeed, degradation classification algorithm in blind image quality indices (BIQI) metric help to describe how distortions affect image statistics and how these distorted statistics can be used to classify images into distortion categories considered here. Recently, Xiongkuo Min et al. (Min *et al.*, 2018) propose distortion-specific metrics based on the pseudo-reference image (PRI) to estimate the probability of three types of distortions: blockiness, sharpness, and noisiness. SVM classifier is trained to identify the degradations. It is worth to mention that the proposed blind PRI-based (BPRI) metric is opinion-unaware and almost training-free excluding for the degradation classification procedure (Min *et al.*, 2018). In (Chetouani *et al.*, 2012), a degradation classification method for natural images is proposed that is based on the recognition accuracy of degradation type and overall image quality assessment. In this paper, each degradation type is considered here as a particular class. Linear discriminant analysis (LDA) is used for the classification of different degradations. A Bayesian approach was utilized in order to predict the type of distortion in images using image quality metrics in (A. Chetouani, 2010, Pages: 714-717). In (Lins *et al.*, 2010), a Randon Forest classifier is used to classify the existence of noise in a given document into six different classes: Back-to-front interference, Frame or border noise, Skew, Orientation (0, 90, 180, 270 degrees), Blur, No noise. In order to classify noise types exist in a document, it is very important to determine which features of a document should be useful. Indeed, a number of features are extracted from each image for classification. The proposed classifier in this work presented a performance standard that is dependable enough to free humans of the burden of selecting which filters to utilize to delete the distortions. Fitri Amia et al. (Arnia *et al.*, 2015) proposed a method to characterize three different types of noises (fox, spots, and uneven background) in historical document based on the discrete cosine transform (DCT) coefficient distribution of the image.

1.2.2 Ancient images

Dedradation of historical documents come from two sources, i.e. noises due to document's age and noises due to digitalization process. Existing classification models for document images, which are degraded due to the digitalization process, were mentioned in the previous section. To the best of our knowledge, no attention has been paid to the classification of physical degradations in historical documents images. In order to maintain, control and enhance the quality of degraded documents and also decrease the negative effect of distortions on diverse processing and analysis systems, it is necessary to estimate the probability of different distortion types in ancient document images.

1.2.3 Discussion

As discussed above, some image degradation classification models are designed for natural images. As mentioned above, there is no attention to propose a degradation classification model for physical distortions of ancient document images in the literature. Proposing a degradation classification model can be useful for many purposes, such as enhancement of degraded documents, optimization of enhancement parameters and etc. A degradation classification model for physical distortions can be a first and important step for designing an automatic degradation modeling that can detect and estimate the type and severity of all existing noises in ancient document images, respectively.

1.3 Evaluation measures for objective quality assessment

There are some evaluation metrics in order to calculate the relation between MOS/DMOS and objective predictions. Some of these metrics are explained in the following. It should be noted that a good objective quality is expected to obtain low values in RMSE and high values in PCC, SRC and KRC.

1.3.1 Pearson's Linear Correlation Coefficient (PLCC)

PLCC is based on a linear regression analysis of pairs of values taken from different data sources. The PLCC produces values that range from -1 to 1. Given a set of MOS values and the respective values predicted by an objective quality metric, denoted as OM, PLCC measure is calculated by:

$$PLCC = \frac{\sum_{i=1}^{L}(MOS_i - \widetilde{MOS})(OM_i - \widetilde{OM})}{\sqrt{\sum_{i=1}^{L}(MOS_i - \widetilde{MOS})^2}\sqrt{\sum_{i=1}^{L}(OM_i - \widetilde{OM})^2}} \tag{1.1}$$

where L is the total number of stimuli in the set, i.e. length of MOS (and OM) used for performance evaluation.

1.3.2 Spearman's Rank Order Correlation Coefficient (SROCC)

SRC is considered for evaluation between subjective and objective scores. It is considered as a measure of prediction monotonicity. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. The SROCC is defined as follows:

$$SROCC = 1 - \frac{6\sum(d_i^2)}{L(L^2-1)} \tag{1.2}$$

where d_i is the difference between the rank of the i-th stimulus in subjective and objective evaluation and L is the total number of stimuli in the set, i.e. length of MOS (and OM) used for performance evaluation.

1.3.3 Kendall's Rank order Correlation Coefficient (KRCC)

KRC are used to assess the similarities among the different rankings. For each pair of rankings, the KRC coefficient is computed to measure the degree of similarity between two sets of rankings obtained by two different techniques on a same set of features (or attributes). The Kendall τ coefficient is defined by:

$$KROCC = \frac{\text{(number of concordant pairs)} - \text{(number of discordant pairs)}}{\frac{1}{2}L(L-1)} \tag{1.3}$$

where L is the total number of stimuli in the set, i.e. length of MOS (and OM) used for performance evaluation. It should be mentioned that if the order in terms of MOS and OM agrees, the pair is considered “concordant”. In the opposite case, the pair is “discordant”.

1.3.4 Root Mean Square Error (RMSE)

RMSE is commonly used in IQA community to measure the prediction accuracy of objective scores. RMSE can be computed by the following formula:

$$\text{RMSE} = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(MOS_i - OM_i)^2} \tag{1.4}$$

where L is the total number of stimuli in the set, i.e. length of MOS (and OM) used for performance evaluation.

CHAPTER 2

OBJECTIVES AND GENERAL METHODOLOGY

In this chapter, we briefly expose our objectives and general methodology. Indeed, we focus on the three main goals of subjective, objective and degradation classification of degraded document images. The developed datasets, proposed objective quality assessment metrics and degradation classification model help to better understand and analyze of document image quality and physical degradation in historical document images. More details and explanations for each objective will be provided in the following chapters (Chapter 3-6). To address particular objectives that were mentioned in the introduction chapter, we set four objectives to be tackled in this work.

2.1 Research objectives

The main goal of this thesis is to develop HVS-based datasets and propose objective image quality assessment metrics and degradation identification model for historical document images. It will be achieved with three specific objectives that all of them related to the human visual perception of document image quality that are summarized as follows:

2.1.1 Objective O1: Subjective and objective quality assessment of historical document images based on human judgments (PS1-PS4 and RQ1)

In the literature, there is no dataset of old and degraded document images consisting of the full-page color image with associated MOS values. Also, there is not any metric for quality evaluation of ancient document images in the literature. Therefore, we were motivated to develop a dataset for degraded document images based on human judgments and to propose a quality evaluation metric. Our approach to developing the VDIQA dataset is detailed in Chapter 3. A no-reference image quality assessment metric is proposed using the analysis of the spatial domain statistics of document images. The proposed metric follows a property of the HVS that evaluates text and non-text regions differently. The experimental results verify the promising performance and efficiency of the proposed metric in comparison with the state-of-the-art NR-IQA metrics on the developed dataset (VDIQA). We describe the proposed metric in Chapter 3.

2.1.2 Objective O2: Subjective and objective quality assessment as well as degradation identification for physical distortions in damaged manuscripts (PS1-PS5 and RQ2)

In order to classify and model different degradation types in ancient document images, we need to develop a dataset with several categories of physical distortions. Therefore, a dataset with four categories of common degradation types in historical document images is developed based on the human perception. We believe that the proposed dataset is the first dataset which has both MOS values and different types of physical degradations for degraded document images (Chapter 4). For the purpose of quality evaluation of the degraded documents, a new blind IQA metric is proposed. According to the performance analysis of the seven no-reference image quality assessment metrics, the proposed metric achieved the highest correlation with the human judgments on two datasets (VDIQA and MHDID). Also, a degradation modeling based on the proposed metric is defined to estimate the probability of different type of degradations in damaged manuscripts. More details and explanations on the proposed metric and degradation classification model are provided in Chapter 5.

2.1.3 Objective O3: Propose a new dataset and a quality assessment metric for degraded medieval document images based on a saliency approach (PS1-PS4 and RQ3)

Color is an important feature in degraded medieval document (DMD) images. For the purpose of assessing the quality of the DMD images, we first created a dataset with MOS values and proposed a quality assessment metric. The proposed metric is based on the analysis of the color saliency in DMD images. Unlike the aforementioned metrics (O1 and O2), this metric is specifically designed to deal with one type of document images (DMD images) with the goal of maximizing the performance. The proposed specific purpose metric will be explained with more details in chapter 6.

2.2 General methodology

We propose three methodologies M1, M2 and M3 to respectively address the research questions RQ1, RQ2 and RQ3 as well as the specific objective O1, O2 and O3. The three methodologies are defined as follows:

2.2.1 Methodology M1: First dataset and efficient blind IQA metric for degraded documents

We introduce a visual document image quality assessment (VDIQA) dataset using PCR for visual quality assessment of degraded document images. There are few DIQA datasets available in literature which are based on the OCR performance. In what follows, we mention why human judgments are used in the introduced dataset instead of the OCR accuracy. Firstly, the OCR engines are not perfect specially for some of the languages and old writing styles and fonts. The second reason is that high OCR accuracy does not necessary mean that a document image is of high quality, rather it means that text region is not degraded. This latter avoids using DIQA for the applications like automatic aging of documents. VDIQA is developed to be used in future researches related to the degraded document image analyses (Chapter 3).

A no-reference image quality assessment metric, which is called VDQAM, is proposed using the analysis of the spatial domain statistics of document images. In this metric, each degraded document image is segmented into four layers based on the log-Gabor filter. This segmentation is based on the assumption that the sensitivity of the human visual system (HVS) is different at the locations of text and non-text. The experimental results demonstrates that the proposed metric has better performance in comparison with the state-of-the-art metrics on the VDIQA dataset (Chapter 3).

2.2.2 Methodology M2: A new blind quality assessment metric, dataset and degradation identification algorithm

It should be mentioned though that the number of images in VDIQA dataset remains small and we dispose with no information about the types of distortion in this dataset, as this information is crucial for subsequent processing. Therefore, we proceeded by creating a new dataset which contains more details about degradation types and also contains more degraded documents in comparison with our previous developed dataset (VDIQA). In this work, we present three contributions that are explained as follows.

For the first contribution, a new database (MHDID) has been developed for the purpose of visual quality evaluation of historical document images. This dataset are classified into four categories based on their distortion types, namely, paper translucency, stain, readers' annotations, and worn holes. The second contribution of this work is the proposition of an efficient Multi-distortion Document Quality Measure (MDQM) for quality evaluation of physically degraded document images. The proposed MDQM metric is based on three sets of spatial and frequency image features. These features are extracted from two layers of text and non-text and mapped to the MOS values using regression function. The third contribution of this work is proposing a degradation classification model to estimate the probability of four common distortion types. Indeed, a multi-class SVM classifier is used to classify the degraded images into four different distortion categories paper translucency, readers' annotations, stain and worn holes in this work (Chapter 4 and Chapter 5).

2.2.3 Methodology M3: Subjective and objective quality assessment metric for degraded medieval documents

In order to propose a blind quality assessment metric for DMD images, we need a DMD dataset to test and train the proposed metric. Therefore, we introduced a new dataset with 150 DMD images which is called DMDD. This dataset is introduced based on the PCR method. Then, we propose a new no-reference metric (BQAD) in order to evaluate the quality of DMD images in the developed dataset (DMDD). The proposed metric is based on the fact that HVS has different sensitivity to the pictorial and textual parts of a DMD. A new color saliency approach and a phase-based binarization method are used for segmentation of the DMD images into three layers. The extracted features from these layers are mapped to the subjective quality scores by regression analysis. It should be mentioned that the developed dataset and the proposed metric are the first attempt to assess the quality of DMD images (Chapter 6).

CHAPTER 3

ARTICLE I- SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT OF DEGRADED DOCUMENT IMAGES

Atena Shahkolaei [1], Hossein Ziaei Nafchi [1], Somaya Al-Maadeed [2], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Department of Computer Science and Engineering, Qatar University, Doha, Qatar

Published in Journal of Cultural Heritage
Volume 30, April 2018, Pages 199-209

Abstract

The huge amount of degraded documents stored in libraries and archives around the world needs automatic procedures of enhancement, classification, transliteration, etc. While high quality images of these documents are in general easy to be captured, the amount of damage these documents contain before imaging is unknown. It is highly desirable to measure the severity of degradation that each document image contains. The degradation assessment can be used in tuning parameters of processing algorithms, selecting the proper algorithm, finding damaged or exceptional documents, among other applications. In this paper, the first dataset of degraded document images along with the human opinion scores for each document image is introduced in order to evaluate the image quality assessment metrics on historical document images. In this research, human judgments on the overall quality of the document image are used instead of the previously used OCR performance. Also, we propose an objective no-reference quality metric based on the statistics of the mean subtracted contrast normalized (MSCN) coefficients computed from segmented layers of each document image. The segmentation into four layers of foreground and background is done on the basis of an analysis of the log-Gabor filters. This segmentation is based on the assumption that the sensitivity of the human visual system (HVS) is different at the locations of text and non-text. Experimental results show that the proposed metric has comparable or better performance than the state-of-the-art metrics, while it has a moderate complexity.

Keywords
Document image quality assessment, degraded document images, physical noises, MSCN coefficients, human visual system, log-Gabor filter.

3.1 Introduction

Heritage is our legacy from the past, what we live with today, and what we pass on to future generations (United Nations Educational, Scientific and Cultural Organization (UNESCO)). Damaged manuscripts and documents constitute an important part of the heritage. They are the memory of human cultures, their history, their achievements, their lifestyle and their individual and social behaviors (Hedjam *et al.*, 2015). Recent years have seen increasing efforts in archiving and digitizing these valuable documents in order to safeguard them against deterioration, and to process them by signal processing techniques (Antonacopoulos & Downton, 2007; Hedjam *et al.*, 2015; Savino & Tonazzini, 2016). Having a large amount of the documents archived, it is very interesting to use automatic techniques to extract and use meaningful information from their images. Because of aging, improper handing and environmental factors, these documents suffer from low to high degrees of degradation. These physical degradations appear on the digitized document images as well. According to the taxonomy of (Lins, 2009), these types of degradation refer to the “physical noises" in document images. The following, is a list of the physical noises (Lins, 2009): folding marks, paper translucency, paper aging, paper texture, paper punching, stains, torn-off regions, worm holes, readers’ annotations, physical blur, carbon copy effect, scratches and cracks, sunburn, and inadequate printing. Physical noises can greatly affect the performance and accuracy of the document processing algorithms.

In order to reduce the effect of the physical noises on the document processing algorithms, it is common to automatically enhance document images. The performance of the current enhancement techniques is not reliable, and the reason can be explained by two arguments.

Firstly, document images might be extremely degraded such that their enhancement is beyond the ability of the current enhancement methods. In addition, the enhancement techniques may over/under-estimate the severity of degradations which leads to an inaccurate enhancement. We note that other imaging techniques like multi-spectral may help (Walvoord & Easton, 2008; Liang, 2012), but it is beyond the scope of the current study.

More specifically, we noted that the performance of the current enhancement methods is far beyond the human visual system's (HVS) ability to distinguish type and severity of degradations. Computational models of HVS that are able to assess common distortions appearing in natural images, synthetic images, etc. are available and evolving (Wang *et al.*, 2004; Sheikh *et al.*, 2006; Mittal *et al.*, 2012; Ghadiyaram & Bovik, 2017; Kundu & Evans, 2015). Several objective image quality assessment (IQA) metrics based on these models have been proposed. Similarly, document images can be assessed and benefit from the quality metrics. Unfortunately, less attention has been paid to the quality assessment of physically degraded document images.

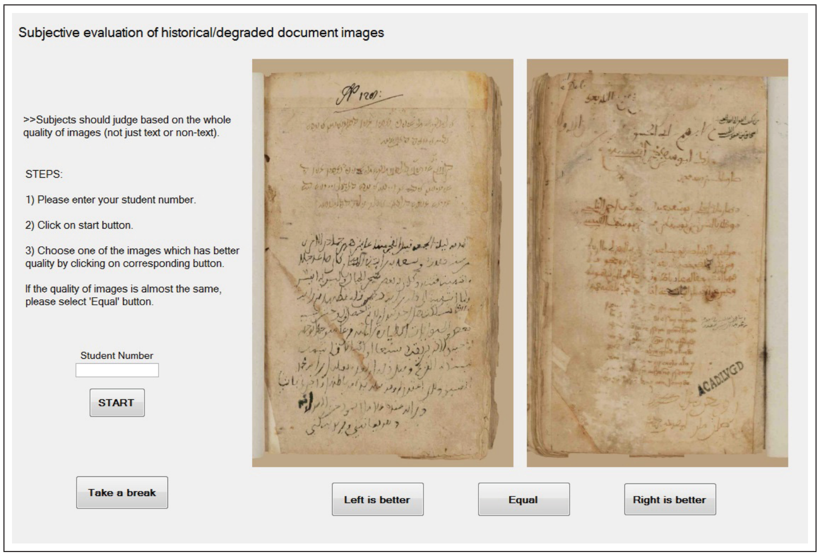
Given the ability of the HVS to recognize degradations in document images, it is very interesting to develop degradation assessment metrics that mimic the HVS. In turn, enhancement methods or other processing algorithms can be developed or tuned in relation to these metrics. Visual document image quality assessment (VDIQA), if not used by the enhancement methods, can have other applications. It helps in the classification of the document images on the basis of their visual quality. For example, an extremely degraded document can be picked and treated manually in order to extract maximum information or to prevent further document deterioration. The performance expectation of the document processing algorithms can be estimated through VDIQA metrics.

The applications of VDIQA and the hypothesis that a document image can be assessed even though it cannot be enhanced, motivated us to develop the first VDIQA dataset based on HVS judgments. There are few document image quality assessment (DIQA) datasets available in the literature, and most of them are evaluated based on the OCR performance. In what follows, we mention why human judgments are used in the introduced dataset instead of OCR accuracy. Firstly, the OCR engines are not perfect, specially for some of the languages, old writing styles and fonts. The second reason is that higher OCR accuracy does not necessarily mean that a document image is of high quality, rather it may mean that the text region is not degraded. This avoids using DIQA for applications like automatic age estimation of documents. In this paper, we use the term VDIQA when human judgments are used and DIQA when OCR accuracy is used. Image quality assessment (IQA) of natural images, mostly focuses on assessment of distortions caused in the acquisition setup, compression, data transfer, etc. However, we suppose that high quality images of documents are available in our study. In other words, document images are free from the distortions caused in the acquisition stage. Of course, other distortions such as blur effect, can be assessed; but the complexity of the problem will increase if other distortions are taken into account.

IQA datasets of natural images (Sheikh *et al.*, 2006; Ponomarenko *et al.*, 2013), synthetic images (Kundu & Evans, 2015), photo retouched images (Vu *et al.*, 2012), and screen content images (Yang *et al.*, 2015) are publicly available in a fast growing field of research. However, there has been little effort to develop datasets and metrics for quality assessment of the document images. The majority of the datasets that have been introduced are either not available or not available to the public (Chou & Yu, 1993; Govindaraju & Srihari, 1995; Kulesh *et al.*, 2001; Obafemi-Ajayi & Agam, 2012). The very few quality metrics for DIQA are either not available or not available to the public (Chou & Yu, 1993; Cannon *et al.*, 1999; Park *et al.*, 2000; Kumar & Ramakrishnan, 2011; Obafemi-Ajayi & Agam, 2012; Kulesh *et al.*, 2001). The authors of (Eilertsen *et al.*, 2013) surveyed DIQA/VDIQA metrics and datasets. The DIQA datasets can have different characteristics. Images in a dataset might be in the form of color, gray-level, or binary. Each image may show a character, a word, a sentence, or a full page. OCR accuracy and mean opinion score (MOS) are two data types available in these datasets for the evaluation purposes. The majority of the VDIQA datasets is in the form of binary images (Chou & Yu, 1993; Govindaraju & Srihari, 1995; Obafemi-Ajayi & Agam, 2012). The dataset introduced in (Kumar *et al.*, 2012) consists of 135 gray-level images with blur distortion. The MOS values for this dataset are collected and computed by crowd-sourcing. To the best of our knowledge, there is no dataset of old and degraded document images consisting of full page color images with associated MOS values. The lack of such a dataset motivated us to develop the first of its kind.

Quality of images can be assessed subjectively and objectively. Subjective evaluation is more accurate and ecologically valid. However, it involves human participation; therefore it is time-consuming and expensive. To overcome these limitations, objective metrics have been proposed in the literature. Objective IQAs are mathematical models that approximate the results of subjective IQA. The main goal of objective IQA is to supply quality metrics that can automatically predict perceived image quality. According to the availability of non-distorted images, objective image quality assessment can be classified into three categories: full reference (FR)(Chandler & Hemami, 2007; Wang *et al.*, 2004; Zhang *et al.*, 2011; Sheikh & Bovik, 2006; Nafchi *et al.*, 2016), reduced-reference (RR)(Wang *et al.*, 2006; Rehman & Wang, 2012; Li & Wang, 2009), and no-reference (NR) IQA models (Moorthy & Bovik, 2011; Saad *et al.*, 2012; Mittal *et al.*, 2012; Liu *et al.*, 2014; Gu *et al.*, 2016). In NR-IQA, only distorted images are available. This paper focuses on historical document images, for which the reference image is not available.

NR-IQA metrics perform according to the statistical regularities of natural images in spatial and transformed domains. The deviation between statistical regularities of distortion-free and distorted images is considered in the design of the NR-IQA models. The so-called NR-IQA metric DIVINE (Moorthy & Bovik, 2011), first classifies distortion types. Then, subband coefficients of discrete wavelet transform (DWT) are fitted by generalized Gaussian distribution (GGD). The statistics of GGD determine the severity of distortions and quality scores are thus estimated by regression. BLIINDS-II (Saad *et al.*, 2012) is a non-distortion specific NR-IQA metric based on the statistics of the discrete cosine transform (DCT) coefficients. The popular NR-IQA metric BRISQUE (Mittal *et al.*, 2012) uses the statistics of natural images in the spatial domain. The distribution of mean subtracted contrast normalized (MSCN) coefficients in two image scales is fitted by symmetric GGD and asymmetric GGD. MSCN coefficients are widely used by NR-IQA models. Similar to the metric DIVINE, CurveletQA (Liu *et al.*, 2014) is also a two stage distortion classification and distortion severity estimation NR-IQA model. It performs according to the statistics of the curvelet coefficients extracted from the images after applying the curvelet transform. BQMS (Gu *et al.*, 2016) is a NR-IQA metric specifically proposed for the quality assessment of screen content images. It performs by using the principles of free energy theory. Recently, an effective NR-IQA metric (FRIQUEE (Ghadiyaram & Bovik, 2017)) was proposed; it is based on a bag of features approach. The metric extracts a large number of features in spatial and frequency domains, and considers color features extracted from different color spaces.



Subjective evaluation of historical/degraded document images

>>Subjects should judge based on the whole quality of images (not just text or non-text).

STEPS:

1) Please enter your student number.

2) Click on start button.

3) Choose one of the images which has better quality by clicking on corresponding button.

If the quality of images is almost the same, please select 'Equal' button.

Figure 3.1 Screenshot of the interface used for subjective evaluation.

None of these metrics are designed for quality assessment of document images. Basically, they do not access the distortions appearing on documents because of aging and physical condition. Several DIQA metrics have been proposed to assess specific distortions such as blur. The first non-distortion specific DIQA metric was proposed on the basis of unsupervised feature learning (Ye & Doermann, 2012). The nonuniform patches corresponding to the text regions are used to construct a visual codebook. The extracted unsupervised features are then mapped to the OCR accuracy by support vector regression (SVR). The problem with this approach is that it relies only on the features that are extracted from the text regions. The performance of this metric is also sensitive to the size of the codebook, which was later improved in (Ye *et al.*, 2013). In (Kang *et al.*, 2014), a quality score is computed for nonuniform patches on the basis of the convolutional neural network (CNN). The average of these quality scores is considered as the quality score for a document image. Xu et al. (Xu *et al.*, 2016) proposes a DIQA metric that uses high order statistics features as well as the local and global features. This metric shows higher performance in comparison with the metrics in (Ye *et al.*, 2013) and (Kang *et al.*, 2014). The above-mentioned non-distortion specific DIQA metrics are tested on the datasets that do not contain many of the types and severity of physical distortions that are common in historical document images. Also, these metrics work on text regions to correlate with OCR accuracy.

Here, we propose a VDIQA metric that take into account distortions in the text and non-text regions of the document images. The proposed metric is tested on document images that contain different types and severity of physical distortions. Appealing to log-Gabor wavelets, we propose a new method to segment document images into four layers of text and non-text. For quality assessment, the extracted features from the analysis of the MSCN statistics of these four layers are mapped to the MOS values by SVR. The experimental results show the promising performance and efficiency of the proposed metric.

The rest of the paper is organized as follows. Section 3.2 provides a full description of the subjective evaluation and creation of the dataset as well as its characteristics and possible applications. The proposed metric for visual document image quality assessment is presented in Section 3.3. Section 3.4 provides the experimental results and discussion, followed by the conclusion in Section 7.3.

3.2 VDIQA dataset

This section describes the procedure for constructing the developed dataset as well as its characteristics. A total of 177 degraded document images with the diverse degradations were selected. The document images are taken from the library of Qatar University. The images are a subset of pages from several old Arabic books. These books belong to the 1^{st} to the 14^{th} Islamic centuries. While the document images contain diverse type and severity of degradation, their format is somewhat similar as is common in Arabic texts. Arabic scripts are used to write different languages, such as Arabic, Urdu, Persian, Pashto, etc. The most important styles that are used for writing Arabic scripts are Nastaliq and Nasth, which read from right to left (Bukhari *et al.*, 2011). Document images in the developed dataset might have colored text, but they do not contain graphics.

3.2.1 Subjective evaluation

For the purpose of visual evaluation of the document images, 30 graduate students whose main field is image processing, electrical engineering, and telecommunications were employed to judge the quality of the document images. The subjects were asked to give their ratings based on the overall quality of the document images. They were all given basic information on what physical noise means in document images. We used a pair comparison ranking (PCR) rating system as in (Sun *et al.*, 2017). At each iteration, two document images on a 17-inch monitor (LCD) at the resolution of 1280×1024 pixels are shown to the subject. The subject had to select the better-quality image. For cases that were not easy to judge, the subjects could rate them as equal. Fig. 3.1 illustrates the rating procedure.

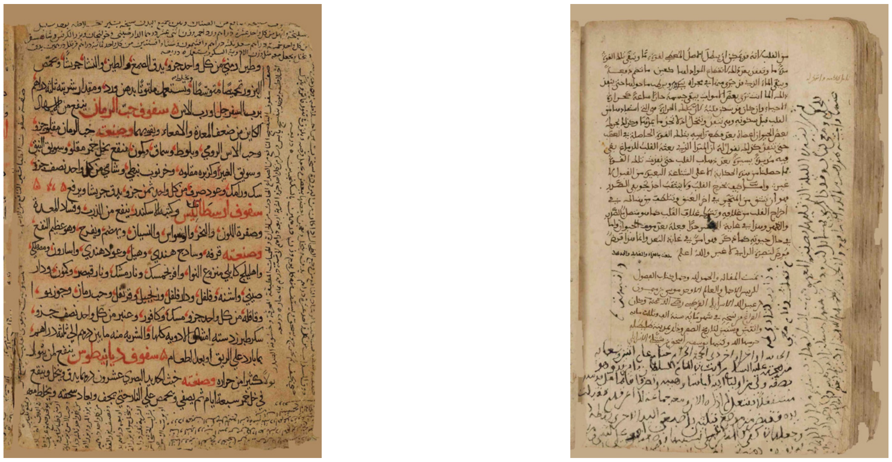
Each pair was compared by at least ten subjects. The average agreement between subjects was 81.34%. We intentionally included a few pages without text in the dataset (Fig. 3.2), while some images have side-notes as well as main-body text (Fig. 3.3). The observation was that subjects had a maximum disagreement rate on these pages. Each subject was told that a page without text does not necessarily mean that its text had been removed as a result of degradation.

We then removed judgments of three subjects as being outliers. As a result, the average agreement increased to 84.72%. The judgments given by the 27 remaining subjects were taken into account and normalized between 0 and 9, where a quality score of 0 means worst quality

(MOS = 4.09)

(MOS = 3.47)

Figure 3.2 Two examples of document images in the VDIQA dataset that contain no text.

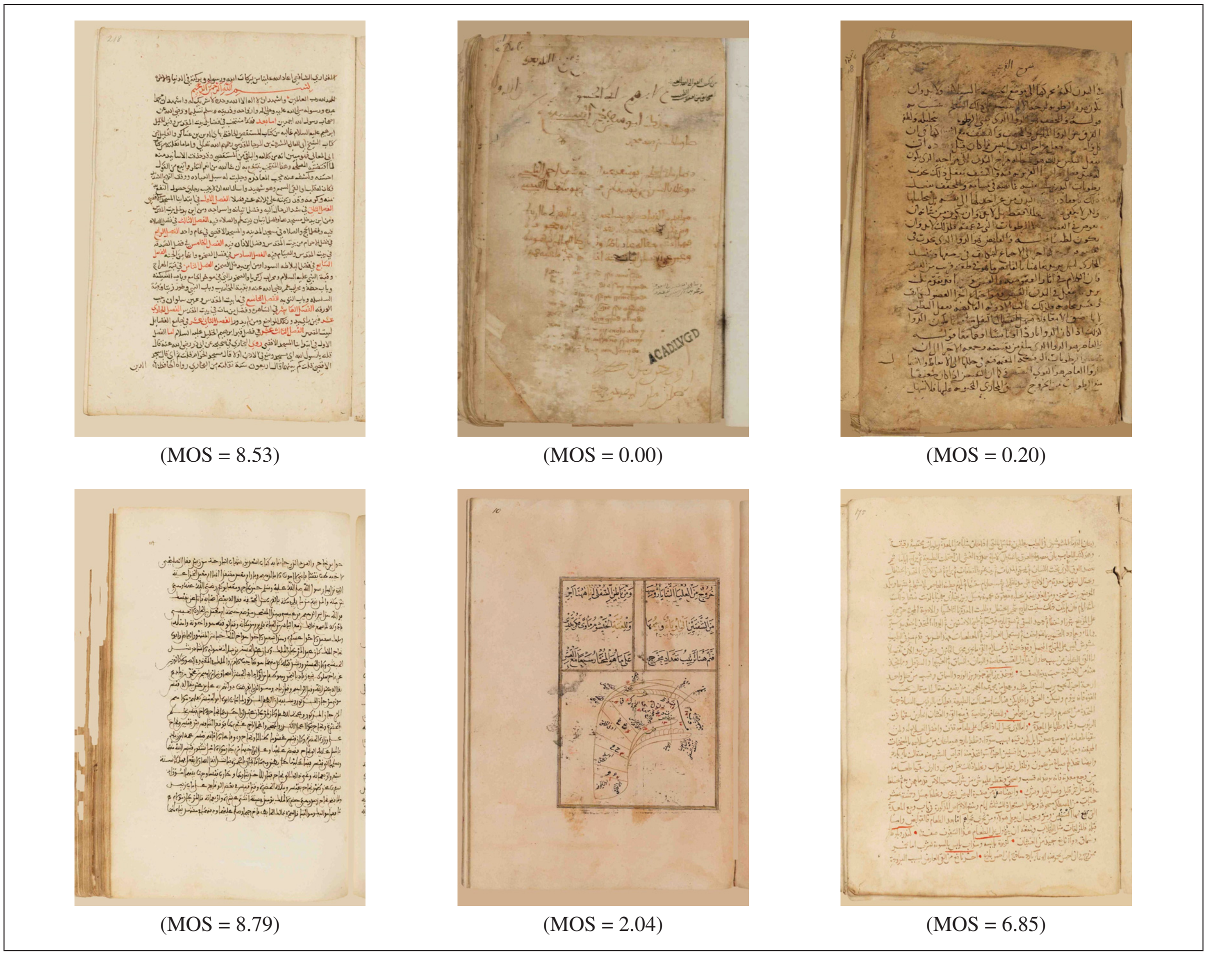


(MOS = 5.62)

(MOS = 6.85)

Figure 3.3 Two examples of document images in the VDIQA dataset that contain side-notes and main-body text.

and 9 indicates best quality. In Fig. 3.4, six ancient document images with various types of degradation are shown along with their MOS value.



(MOS = 8.53) (MOS = 0.00) (MOS = 0.20)

(MOS = 8.79) (MOS = 2.04) (MOS = 6.85)

Figure 3.4 Six examples of the degraded document images in the developed dataset with their MOS values. Higher MOS values indicate better-quality images.

The histogram of MOS values in the VDIQA dataset is shown in Fig. 3.5. As can be seen, the MOS values are fairly distributed for highly degraded, moderately degraded, and lightly degraded document images in the VDIQA dataset.

Apart from the degradation type and severity, it is common to show that images in a dataset contain diverse original content from the edge and color perspective. To this end, spatial information (SI) and colorfulness (CF) as suggested in (Winkler, 2012) were used in this study. For each image in the dataset, SI indicates the edge energy, and CF is a measure of color richness. Fig. 3.6 shows the scatter plot of SI versus CF for 177 images in the dataset. As can be seen, unlike the CF interval, which is relatively small, the SI interval is large. This conclusion is

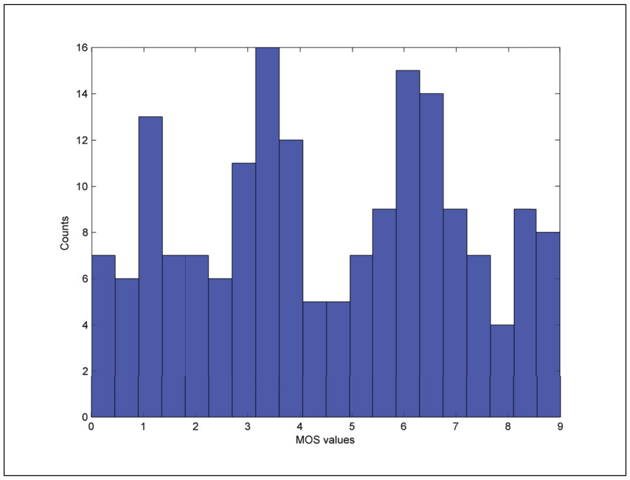


Figure 3.5 Histogram of MOS values in the developed dataset. Lower MOS values indicate lower-quality document images.

based on a comparison with the natural datasets (not shown here). It is simply the result of the obvious fact that more colors are used in natural images than the old document images. This is at least valid for the developed dataset in this paper. Also, some low values of SI in Fig. 3.6 indicate pages without text, as shown in Fig. 3.2.

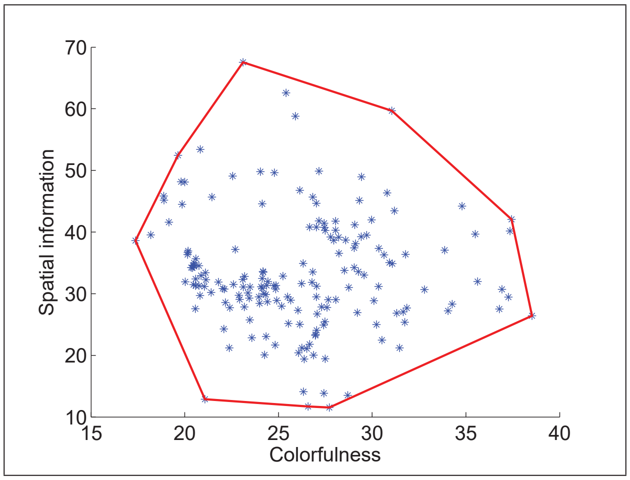


Figure 3.6 A scatter plot of spatial information (SI) against colorfulness (CF) for 177 document images in the developed dataset.

The detailed characteristics of the developed dataset are listed as follows.

- Distortions: Each image contains at least one or more types of the following physical noises common in historical document images: Paper deterioration, bleed-through, show-through, alien ink, ink-smear, ink-noise, faded ink. Images contain only degradations caused by aging. The images contain no distortion as a result of compression, transmission, acquisition setup, etc.
- Number of images: 177 document images from similar books are selected.
- Screen resolution: The same screen size and resolution was used, but manufacturers varied.
- Illumination: Normal office light at our laboratory.
- Number of subjects: 30 subjects, but 3 were excluded in the data processing step.
- Number of ratings per image: At least 10 ratings per each pair.
- Data format: Mean opinion score (MOS).
- Range of scores: MOS values normalized between 0 and 9.
- Image dimension: Size of images is 1400×2000 pixels.
- Image format: The format of images is JPEG and the total size of the dataset is $\sim$ 58MB. Images are in the form of the three channel color (RGB).
- Availability: The dataset is available to the public through the following link: https://dl.dropboxusercontent.com/u/ The dataset can be obtained by contacting the authors as well.

3.3 Proposed VDIQA metric

We propose a quality metric for document images that contain physical noises. The proposed metric takes into account several common attributes of the degraded documents that can also affect human judgments on their quality. First, text and non-text parts of a document image

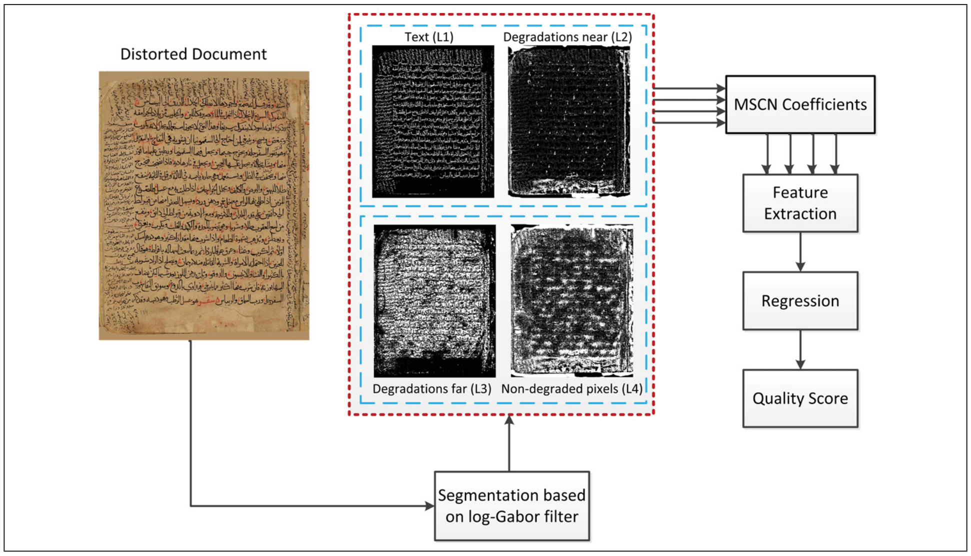


Figure 3.7 Illustration of the proposed metric.

do not have equal impact on HVS judgments. The second consideration is that physical noises close to the text, and those far from the text may not equally contribute to the visual quality of the document images. Finally, the proposed metric supposes that documents with diverse types of degradations are likely to have lower visual quality.

Fig. 3.7 illustrates the proposed metric. The input document image is roughly segmented into four layers of text, possible degradations close to the text, possible degradations far from the text, and possible non-degraded pixels. The statistics of the mean subtracted contrast normalized (MSCN) coefficients (Daniel L Ruderman, 1994) corresponding to each layer are independently computed to form feature vectors. Support vector regression (ε-SVR) is used to map these feature vectors to the MOS values. The following subsections provide further explanations for each part of the proposed metric.

3.3.1 Segmentation based on log-Gabor filter

The first step of the proposed metric is segmentation of the document image into a text layer and three non-text layers. Here, we propose using the log-Gabor filters (Field, 1987) to classify document images because of their ability to simultaneously localize spatial and frequency information. Let $M^e_{\rho r}$ and $M^o_{\rho r}$ denote the even-symmetric and odd-symmetric wavelets at a scale ρ and orientation r which are known as quadratic pairs (Papari & Petkov, 2011). By considering $I(x)$ as a two-dimensional signal on the two dimensional domain of x, the response of each quadratic pair of filters at each image point x forms a response vector by convolving with $I(x)$:

$$[e_{\rho r}(x), o_{\rho r}(x)] = [I(x) * M^e_{\rho r}, I(x) * M^o_{\rho r}] \tag{3.1}$$

where * denotes convolution, and values $e_{\rho r}(x)$ and $o_{\rho r}(x)$ are the real and imaginary parts in the complex-valued frequency domain. We denote the summation of even filter convolution responses over ρ scales and r orientations by the following energy function:

$$\text{E(x)} = \sum_{\rho r} e_{\rho r}(x) \tag{3.2}$$

Since the size of the text and degradations vary from one document image to another, we use 5 filter scales in the experiments. In addition, only one filter orientation (0 degree) was used in the experiments so as to reduce the computation time. Let $\mu_1 < 0$ denotes the mean of the negative values in E(x), and $\mu_2 > 0$ denotes the mean values of the positive values in E(x). The input document image is segmented into four layers by the following thresholding:

$$\begin{cases} L_1(x) = \ \text{E(x)} \leq \mu_1 \\ L_2(x) = \ \text{E(x)} > \mu_1 \ \& \ \text{E(x)} \leq 0 \\ L_3(x) = \ \text{E(x)} > 0 \ \& \ \text{E(x)} < \mu_2 \\ L_4(x) = \ \text{E(x)} \geq \mu_2 \end{cases} \tag{3.3}$$

where, L_1 layer is an approximation of the text regions, while the L_2 to L_4 layers refer to the possible non-text regions of a document image. L_2 may refer to the more severe degradations in a document image, while L_3 refers to the possible degradations around text, and L_4 refers to

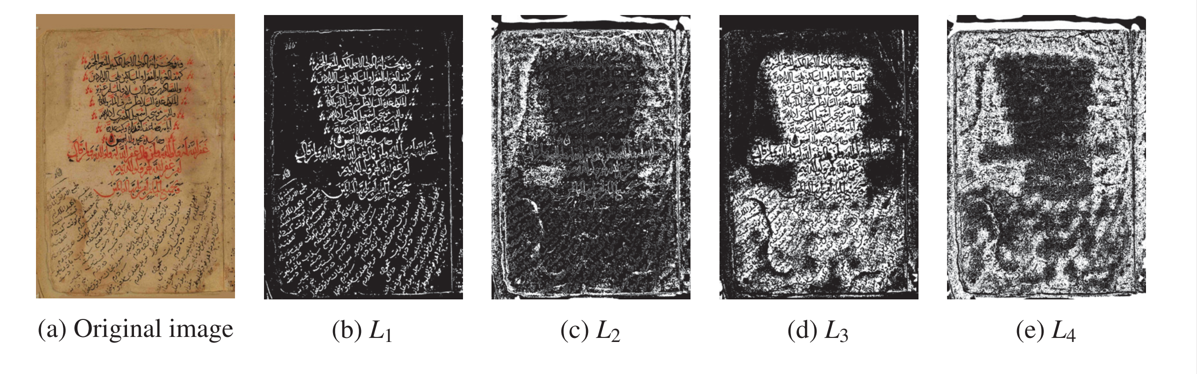
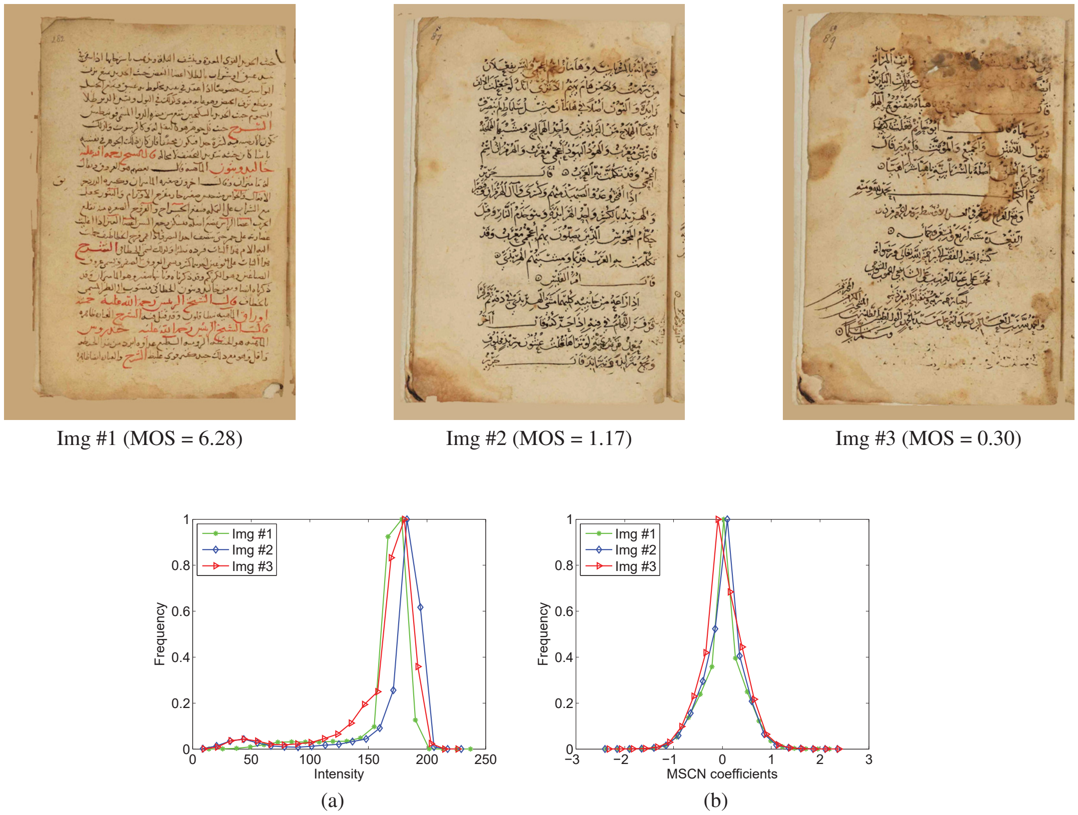

(a) Original image (b) L_1 (c) L_2 (d) L_3 (e) L_4

Figure 3.8 An example of the proposed document image segmentation method into four layers of text and non-text: (a) original image, (b) possible text pixels, (c) possible degradations far from the text, (d) possible degradations close to the text, and (e) possible non-degraded background pixels.

the lightly-distorted or non-distorted pixels. Fig. 3.8 shows a degraded document image along with its four segmented layers. It can be seen that the proposed segmentation method provides a good estimation of the text and non-text layers/pixels. The next step of the proposed metric is to analyze the statistics of the document image pixels at the location of these layers.

3.3.2 Spatial domain document statistics

Due to the different nature and variety of features that degraded document images may have, it is of much interest to use decorrelation techniques that can remove the dependency of these features. Mean subtracted contrast normalized (MSCN) coefficients have been successfully used for quality assessment of natural images (Mittal *et al.*, 2012; Ghadiyaram & Bovik, 2017). MSCN coefficients of a noise-free image should follow a Gaussian distribution. Distorted images will have different shapes of distribution, that can be measured and used for assessment. However, the statistics of the MSCN coefficients of a document image may not reflect the condition of its physical degradations given that a high-quality image of the document is on hand. Therefore, in the proposed metric, statistics of the MSCN coefficients related to the individual segmented layers of the document image are computed. In the following, more details on the use of the MSCN coefficients along with the proposed metric are provided.



Img #1 (MOS = 6.28)

Img #2 (MOS = 1.17)

Img #3 (MOS = 0.30)

(a)

(b)

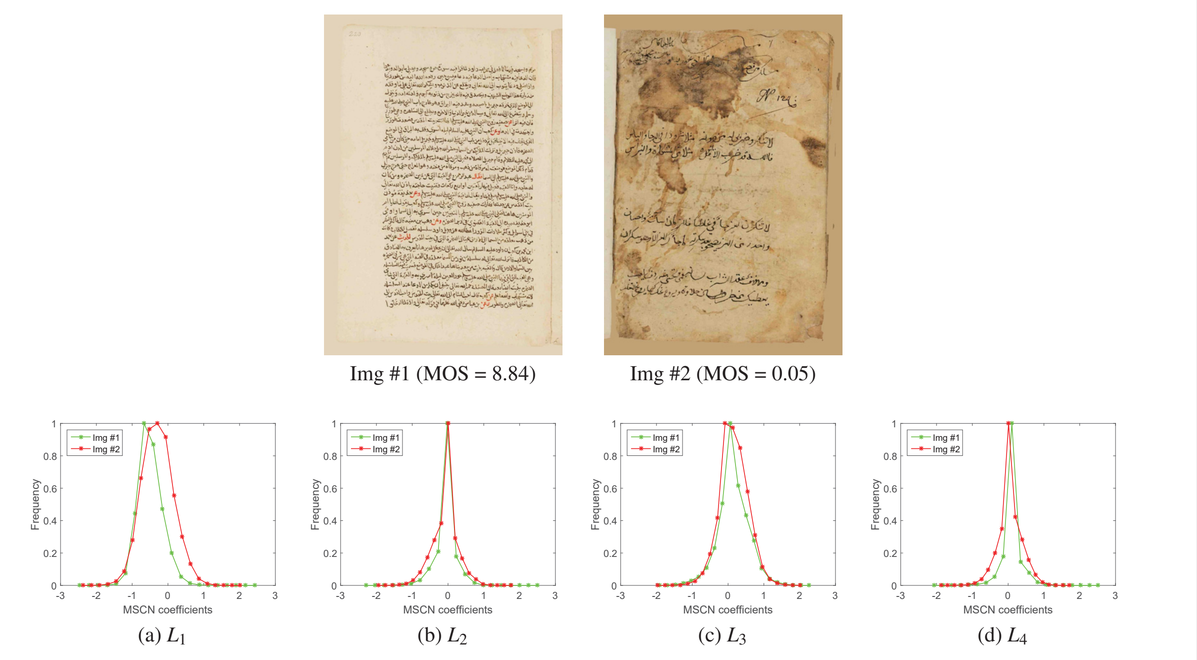
Figure 3.9 Three document images from the developed dataset along with their intensity histogram, and MSCN coefficients histogram.

In (Daniel L Ruderman, 1994), it is shown that applying a local nonlinear operation to a luminance channel of images, using it to remove local mean displacements of luminance, and to normalize local variances of luminance will have a decorrelation effect. Given a 2D input image I, its MSCN coefficients can be computed by:

$$\mathrm{MSCN}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{3.4}$$

where, $C = 1$ is a constant to avoid division by zero, and $\mu(i,j)$ and $\sigma(i,j)$ are defined as

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} I_{k,l}(i,j) \tag{3.5}$$


Img #1 (MOS = 8.84)

Img #2 (MOS = 0.05)

(a) L_1

(b) L_2

(c) L_3

(d) L_4

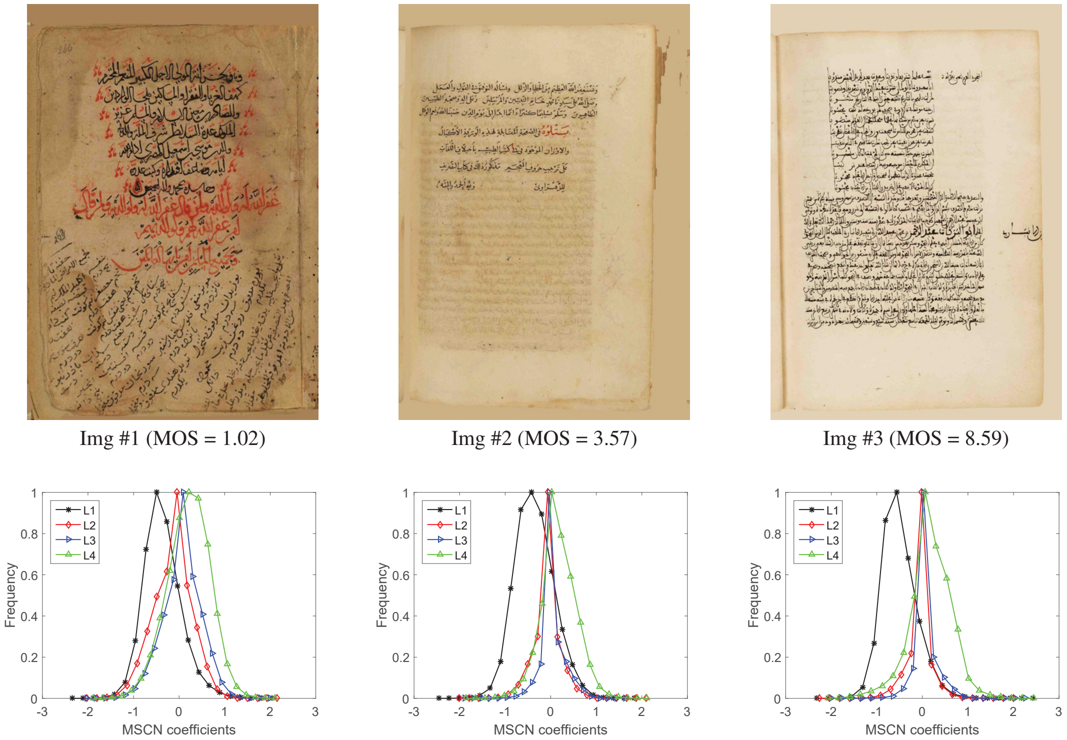
Figure 3.10 Histograms of MSCN coefficients for (a) possible text pixels, (b) possible degraded pixels far from the text, (c) possible degraded pixels close to the text, and (d) possible non-degraded background pixels.

$$\sigma^2(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} \left(I_{k,l}(i,j) - \mu(i,j) \right)^2 \tag{3.6}$$

where, ω is a 2D symmetric Gaussian weighting function. Fig. 3.9 shows examples of the intensity and MSCN distributions for three document images. From the plots in Fig. 3.9, it can be seen that the MSCN coefficients follow a Gaussian distribution, however, the MSCN distributions do not change very much with respect to the physical degradations.

In Fig. 3.10, the MSCN distributions of two document images corresponding to the four segmented layers are shown. The observation is that the MSCN distributions remain Gaussian and that for each layer, the shape of the distributions is different. However, we have shown that the MSCN distributions of the whole document images are very similar (Fig. 3.9).

Finally, Fig. 3.11 shows the joint MSCN histograms of the four layers for three document images. The plots show that the MSCN distributions for the layers of a document image are



Img #1 (MOS = 1.02)

Img #2 (MOS = 3.57)

Img #3 (MOS = 8.59)

Figure 3.11 Histograms of MSCN coefficients for four layers of the three document images.

different which means that the statistics of each layer can contribute to the overall quality score predicted by the proposed metric.

The proposed metric extracts statistical features from each layer of document images in two image scales. The symmetric and asymmetric behavior of the distributions strongly suggests the use of the asymmetric Gaussian distribution analysis in addition to the symmetric one.

3.3.2.1 Symmetric generalized Gaussian distribution

MSCN coefficients are fitted with the generalized Gaussian distribution. The zero mean generalized Gaussian distribution (GGD) is defined by (Sharifi & Leon-Garcia, 1995):

$$f(x;\alpha,\sigma^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} e^{\left(\left(-\frac{|x|}{\beta}\right)^{\alpha}\right)} \tag{3.7}$$

where, σ^2 is the variance of the distribution, and parameter β and gamma function Γ are defined as:

$$\beta = \sigma \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \tag{3.8}$$

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \ , \qquad \alpha > 0 \tag{3.9}$$

where, α controls the shape (and tailness) of the distribution. α and σ are first and second elements of the feature vector. Each degraded document image is segmented into four layers. Therefore, sixteen features are extracted from each image in two scales based on the symmetric generalized Gaussian distribution.

3.3.2.2 Asymmetric generalized Gaussian distribution

When the shape of the MSCN histogram is not symmetric, fitting with generalized Gaussian distribution is not accurate. In this case, asymmetric Gaussian distribution can be used. For the left, and the right side of the distribution, the parameters are estimated independently. Because of the asymmetric behavior of the MSCN distributions, more features are added to the feature vector by estimating the parameters of the asymmetric generalized Gaussian distribution (AGGD). The AGGD with zero mode (Mittal *et al.*, 2012) is defined by:

$$f(x;v,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{v}{(\beta_l+\beta_r)\Gamma(1/v)} \exp\Big(-(\frac{-x}{\beta_l})^v\Big), & x < 0 \\ \frac{v}{(\beta_l+\beta_r)\Gamma(1/v)} \exp\Big(-(\frac{x}{\beta_r})^v\Big), & x \geq 0 \end{cases} \tag{3.10}$$

where, $\beta_l = \sigma_l \sqrt{\Gamma(\frac{1}{v})/\Gamma(\frac{3}{v})}$, and $\beta_r = \sigma_r \sqrt{\Gamma(\frac{1}{v})/\Gamma(\frac{3}{v})}$. The parameter v, which controls the shape (and tailness) of the distribution, σ_l^2 is the variance of the left side of the distribution, and σ_r^2 is the variance of the right side of the distribution. The parameters of the ADDG (v, σ_l^2,

σ_r^2) are estimated based on the proposed method in (Lasmar *et al.*, 2009). The total number of the features extracted from each document image is 40 (20 features for each image scale).

3.4 Experimental results and discussion

In this section, the performance of the proposed visual document quality assessment metric (VDQAM) is analyzed in terms of its ability to predict subjective ratings of image quality. For objective evaluation, three popular evaluation metrics were used in the experiments: the Spearman Rank-order Correlation coefficient (SRC), the Pearson linear Correlation Coefficient (PCC) after a nonlinear regression analysis (equation 6.11), and the Kendall Rank Correlation coefficient (KRC). The SRC and PCC metrics measure prediction monotonicity, and prediction linearity, respectively. The KRC is used to evaluate the degree of similarity between quality scores and MOS. The reported PCC values in this paper were computed after mapping quality scores to MOS according to the following logistic function (Sheikh *et al.*, 2006):

$$f(x) = \beta_1\left(\frac{1}{2} - \frac{1}{1+e^{\beta_2(x-\beta_3)}}\right) + \beta_4 x + \beta_5 \tag{3.11}$$

where β_1, β_2, β_3, β_4 and β_5 are fitting parameters computed by minimizing the mean square error between quality predictions x and subjective scores MOS. For the purpose of visualizing quality scores of the proposed index, the scatter plots of the proposed IQA model VDQAM are shown in Figure 3.12 . The logistic function suggested above is used to fit a curve on each plot.

3.4.1 Performance comparison

The performance of six NR-IQA models and the proposed metric VDQAM on the VDIQA dataset are listed in Table 3.1. The VDIQA dataset is divided into different randomly chosen subsets and the results are reported on the basis of the median value of 1000 times train-test for three cases: 20% train 80% test, 50% train 50% test, and 80% train 20% test. For each evaluation metric in Table 3.1, the top two IQA models are highlighted. The results

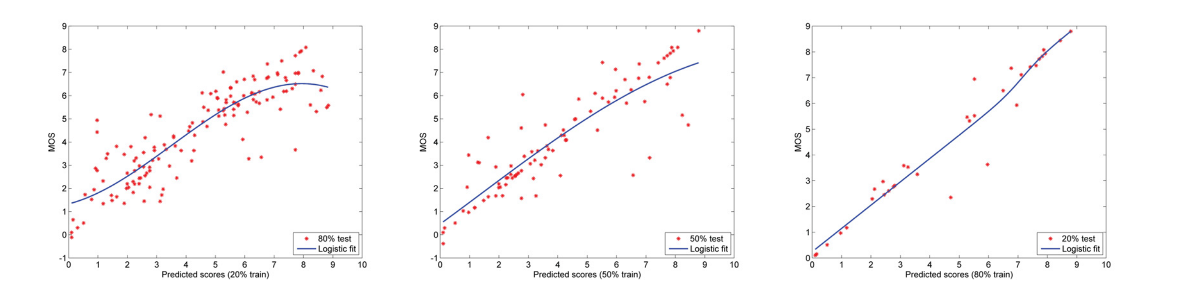


Figure 3.12 Scatter plots of quality scores against the subjective MOS on the VDIQA dataset for the proposed model VDQAM. These scatters are shown based on the median value of 1000 times train-test for three cases: 20% train 80% test, 50% train 50% test, and 80% train 20% test.

Table 3.1 Performance comparison of the proposed metric (VDQAM) and six blind image quality assessment metrics on the developed dataset. Note that top two metrics are highlighted.

NR Indices	SRC	PLC	KRC
20%-80%			
DIVINE	0.6244	0.6459	0.4443
BLIINDS-II	0.5815	0.6025	0.4201
BRISQUE	0.5663	0.5977	0.4019
CurveletQA	0.5031	0.5516	0.3601
BQMS	**0.6539**	**0.6771**	**0.4763**
FRIQUEE	0.5685	0.6326	0.3746
VDQAM	**07008**	**0.7167**	**0.5146**
50%-50%			
DIVINE	0.7313	0.7496	0.5354
BLIINDS-II	0.6826	0.7008	0.5090
BRISQUE	0.7350	0.7574	0.5472
CurveletQA	0.6033	0.6410	0.4435
BQMS	**0.7570**	**0.7801**	**0.5641**
FRIQUEE	0.6996	0.7586	0.4933
VDQAM	**0.7954**	**0.8083**	**0.6189**
80%-20%			
DIVINE	0.7667	0.8048	0.5762
BLIINDS-II	0.7186	0.7591	0.5539
BRISQUE	**0.7810**	0.8084	**0.6076**
CurveletQA	0.6486	0.6990	0.4863
BQMS	0.7737	0.8156	0.5884
FRIQUEE	0.7533	**0.8158**	0.5559
VDQAM	**0.8180**	**0.8493**	**0.6535**

show the promising performance of the proposed metric over the state-of-the-art metrics. The performance of the proposed metric is superior to the popular spatial domain metric BRISQUE thanks to the analysis of the segmented document instead of the raw document image pixels.

In Table 3.2, the correlation between each layer of distorted document images in VDIQA dataset and MOS is reported. The reported results are based on the median value of 1000 times train-test for three cases of 20% train 80% test, 50% train 50% test, and 80% train 20% test. The results in Table 3.2 show that possible degradations close to the text, e.g. layer L_3, achieve a higher correlation than the other layers. This suggests that HVS is more sensitive to the degradations close to the text according to the proposed metric.

Table 3.2 Performance of the individual layers when considered as separate quality assessment models.

		L_1	L_2	L_3	L_4
20%-80%	SRC	0.5283	0.3384	**0.6165**	0.4917
	PCC	0.5513	0.4077	**0.6433**	0.5273
	KRC	0.3681	0.2331	**0.4449**	0.3490
50%-50%	SRC	0.5788	0.4554	**0.6779**	0.5835
	PCC	0.6165	0.4996	**0.7054**	0.6100
	KRC	0.4118	0.3198	**0.4978**	0.4233
80%-20%	SRC	0.6270	0.4867	**0.6886**	0.6369
	PCC	0.6633	0.5582	**0.7375**	0.6770
	KRC	0.4496	0.3659	**0.5116**	0.4658

3.4.2 Computational Analysis

Table 3.3 lists the run times of the seven NR-IQA (NR-VDIQA) models when applied to the images of size 1400×2000. The experiments were performed on a Core i7 3.40GHz CPU with 16 GB of RAM. The IQA models were implemented in MATLAB 2013b running on Windows 7. In addition, the number of features used by each metric is listed in Table 3.3. Efficiency is an important factor for document quality assessment metrics since a large volume of documents is stored in libraries and archives around the world. The proposed metric ranks second fastest among other metrics with a moderate run-time.

Table 3.3 Comparison of six NR-IQA metrics in terms of run time and number of features.

IQA model	no. of features	run time (ms)
BRISQUE (Mittal *et al.*, 2012)	36	823
▷ VDQAM	40	2268
CurveletQA (Liu *et al.*, 2014)	12	17827
DIVINE (Moorthy & Bovik, 2011)	88	145909
FRIQUEE (Ghadiyaram & Bovik, 2017)	560	312230
BQMS (Gu *et al.*, 2016)	13	489910
BLIINDS-II (Saad *et al.*, 2012)	24	522867

3.5 Applications

For the purpose of achieving maximum benefit from the developed VDIQA dataset, a set of possible applications is suggested.

Model selection: Some of the document processing systems are not robust under degradation, while others are designed to have higher performance under the same condition. With an objective measurement of the document image quality, it is possible to choose the proper model on basis of the level of degradation.

Parameter optimization and tuning: the VDIQA metric can help in adaptive selection of parameters for document processing algorithms based on document image quality.

Performance expectation: Errors such as wrong or meaningless information may propagate in a document processing system if they are not predicted.

Educational purposes: VDIQA can be used for educational purposes like handwritten teaching and document repair (Kulesh *et al.*, 2001).

Psychological study: the human visual system (HVS) is very complex (Wang *et al.*, 2004), and models of the HVS can benefit from VDIQA dataset and related studies.

3.6 Conclusion

This paper introduces a dataset for visual quality assessment of degraded document images. Subjective evaluation is conducted on the basis of the ratings of the human visual system on the quality of document images. The developed dataset is a step toward further research on this problem. We have tested the state-of-the-art no-reference image quality assessment metrics on the developed dataset. Also, a no-reference image quality assessment metric is proposed using the analysis of the spatial domain statistics of document images. In the proposed metric, each degraded document image is segmented into four layers according to the log-Gabor filters, and the spatial statistics of these layers are used for quality assessment. The experimental results show the promising performance and efficiency of the proposed metric in comparison with the state-of-the-art NR-IQA metrics on the developed dataset. We believe that the developed dataset and metric help in better understanding and analysis of document image quality and physical degradation.

Acknowledgments

This publication was made possible by NPRP grant NPRP 7-442-1-082 from the Qatar National Research Fund (a member of Qatar Foundation). The statements are solely the responsibility of the authors.

CHAPTER 4

ARTICLE II- MHDID: A MULTI-DISTORTION HISTORICAL DOCUMENT IMAGE DATABASE

Atena Shahkolaei[1], Azeddine Beghdadi[2], Somaya Al-maadeed[3], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Department of Computer Science and Engineering, Paris 13 University, Paris, France

[3] Department of Computer Science and Engineering, Qatar University, Doha, Qatar

Published in IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), March 2018, Pages 156-160

Abstract

In this paper, a new dataset, called Multi-distortion Historical Document Image Database (MHDID), to be used for the research on quality assessment of degraded documents and degradation classification is proposed. The MHDID dataset contains 335 historical document images which are classified into four categories based on their distortion types, namely, paper translucency, stain, readers' annotations and worn holes. A total of 36 subjects participated to judge the quality of ancient document images. Pair comparison rating (PCR) is utilized as a subjective rating method for evaluating the visual quality of degraded document images. For each distortion image a mean opinion score (MOS) value is computed. This dataset could be used for evaluating the image quality assessment (IQA) measures as well as in the design of new metrics.

Keywords

Historical document images, pair comparison rating, physical noises, human visual system, image quality assessment.

4.1 Introduction

Quality assessment and enhancement of historical documents and manuscripts have received a great amount of interest in recent years. Libraries and museum archives are interested in processing and conserving a large number of historical documents using imaging systems. The reason is to avoid time-consuming manual processing and also the fact that this valuable collection of historical documents cannot be preserved forever. A crucial problem in the analysis of historical documents is that they suffer from various degradation types which makes their automatic analysis a difficult task. Creating a new dataset which contains different types of degradations and MOS values of degraded images is the momentous part of the quality evaluation of damaged manuscripts.

Each image may indicate a character, a word, a sentence, or a full page. OCR accuracy and MOS values are two data types available in these datasets for the evaluation purposes (Shahkolaei *et al.*, November 2017). Standard IQA datasets are available for natural images (Winkler, 2012), screen content images (Yang *et al.*, 2015), Arabic cheque databases (Al-Ohali *et al.*, 2003) and photo-retouched images (Vu *et al.*, 2012) in the state-of-the-art. Unfortunately, there are few document image quality assessment datasets in the literature, and these few datasets are not available to the public. Furthermore, these datasets are mostly evaluated according to the OCR performance. In spite of the fact that the majority of visual document image quality assessment datasets are in binary images (Obafemi-Ajayi & Agam, 2012; Chou & Yu, 1993), in some datasets the images are in color or gray-level. Kumar et al. released a publicly available dataset for DIQA (Kumar. *et al.*, 2013). This dataset consists of the camera-captured document images with varying levels of defocus blur introduced manually during the image capture with different cameras. This database provides also the OCR results obtained using some well-known OCR techniques. The dataset introduced in (Kumar *et al.*, 2012) consists of 135 gray-level images with blur distortion. The MOS values for this dataset are collected and computed by crowd sourcing.

Character level OCR accuracy for each image are obtained using several different OCR softwares (Eilertsen *et al.*, 2013). OCR engines have the low performance for some languages, fonts and writing styles. It is worth noticing that, OCR high accuracy is not always associated with high perceptual quality but rather to non-degraded text regions. Therefore, using human-based evaluation instead of OCR performance is more suitable especially for historical documents. To the best of our knowledge, the VDIQA dataset (Shahkolaei *et al.*, November 2017), proposed recently, is the only available dataset for degraded historical documents. This dataset contains 177 full-page color images and their respective MOS values. However, the number of images in this dataset is not sufficient. Moreover, no information about the types of distortions is provided for this dataset.

Image quality evaluation of such databases can be assessed subjectively or objectively. Subjective evaluation is time consuming , cumbersome, expensive and could not be used in real-time applications. . Therefore, objective evaluation is proposed in order to overcome these limitations. Objective evaluation can be categorized into three groups based on the availability of the reference image (Eilertsen *et al.*, 2013): full reference (FR), reduced-reference (RR) and no-reference (NR). Among the different types of objective evaluation categories, NR approaches have attracted more attention, due to the fact that NR IQA metrics do not require the reference image and in most practical cases the reference image is unavailable.

In this paper, a new dataset (MHDID) for degraded document images with four types of degradation classes based on pair comparison rating method is proposed. The same protocol as in (Shahkolaei *et al.*, November 2017) and (Sun *et al.*, 2017) is used to build the MHDID dataset. To the best of our knowledge, the proposed dataset is the first dataset which has both MOS value and different types of distortions for ancient document images.

In the following we provide a detailed description of the MHDID dataset and discuss some open problems and perspectives.

4.2 MHDID dataset description

In this section, we explain the procedure for building the MHDID dataset from historical document images containing various physical degradations that could be classified into different categories.

4.2.1 Selection of distorted images

We selected 335 degraded document images with various degradation types from the library of Qatar University. The total number of ancient document images in this library is around 7000 degraded documents with has at least two types of degradations per image. The language of these manuscripts is Arabic and these documents were collected from 130 different books edited during the 1st to the 14th Islamic centuries period.

The most four common distortion types in this database are the following: paper translucency (88 images), stains (113 images), readers' annotations (61 images) and worn holes (73 images) (Lins, 2009). These degradation types are listed as follows.

- Paper translucency: it is an internal degradation and appears when a document is written on both sides of translucent paper.
- Stains: this degradation is one type of external noises that may occur during the printing process, etc.
- Readers' annotation: sometimes readers make notes and highlight sentences in a document for various reasons. In this case, readers' annotation noise appears.
- Worm holes: it is the dig tunnels in old documents and is easily recognizable by the human visual system (HVS).

Fig. 4.1 illustrates different examples of historical document images with different degradation types selected from four degradation classes of MHDID dataset. The images in each raw correspond to a type of degradation. It can be seen that each degraded image has just a prevailing distortion type which is easily recognized by human perceptual observation.

4.2.2 Subjective image quality evaluation

Subjective image quality assessment methods could be roughly classified into two categories: direct scaling methods and indirect scaling methods as suggested in (Krasula, 2017).

4.2.2.1 Direct scaling methods

Direct scaling methods are based on the judgments of subjects which obtain from each degraded image directly. In the following, three popular direct scaling methods will be explained.

- Absolute category rating (ACR): is considered as the simplest subjective method. In this method, subjects are asked to select one of the five grade scale. These scales are sorted by quality in decreasing order: excellent, good, fair, poor and bad.
- Degradation category rating (DCR): the main purpose of this subjective method is to evaluate how much does the distortion affect the perceived quality. The subjects are asked to select one of the five levels of degradation in each degraded image: imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying.
- Double stimulus continuous quality scale (DSCQC): the reference and the degraded images are displayed to subjects. Then, similarly to ACR and DCR, subjects are asked to categorize images into five different quality scores.

4.2.2.2 Indirect scaling methods

are the sequential-task, more comfortable and less demanding for the subjects in comparison with direct scaling methods. Two common indirect scaling methods will be explained as follows.

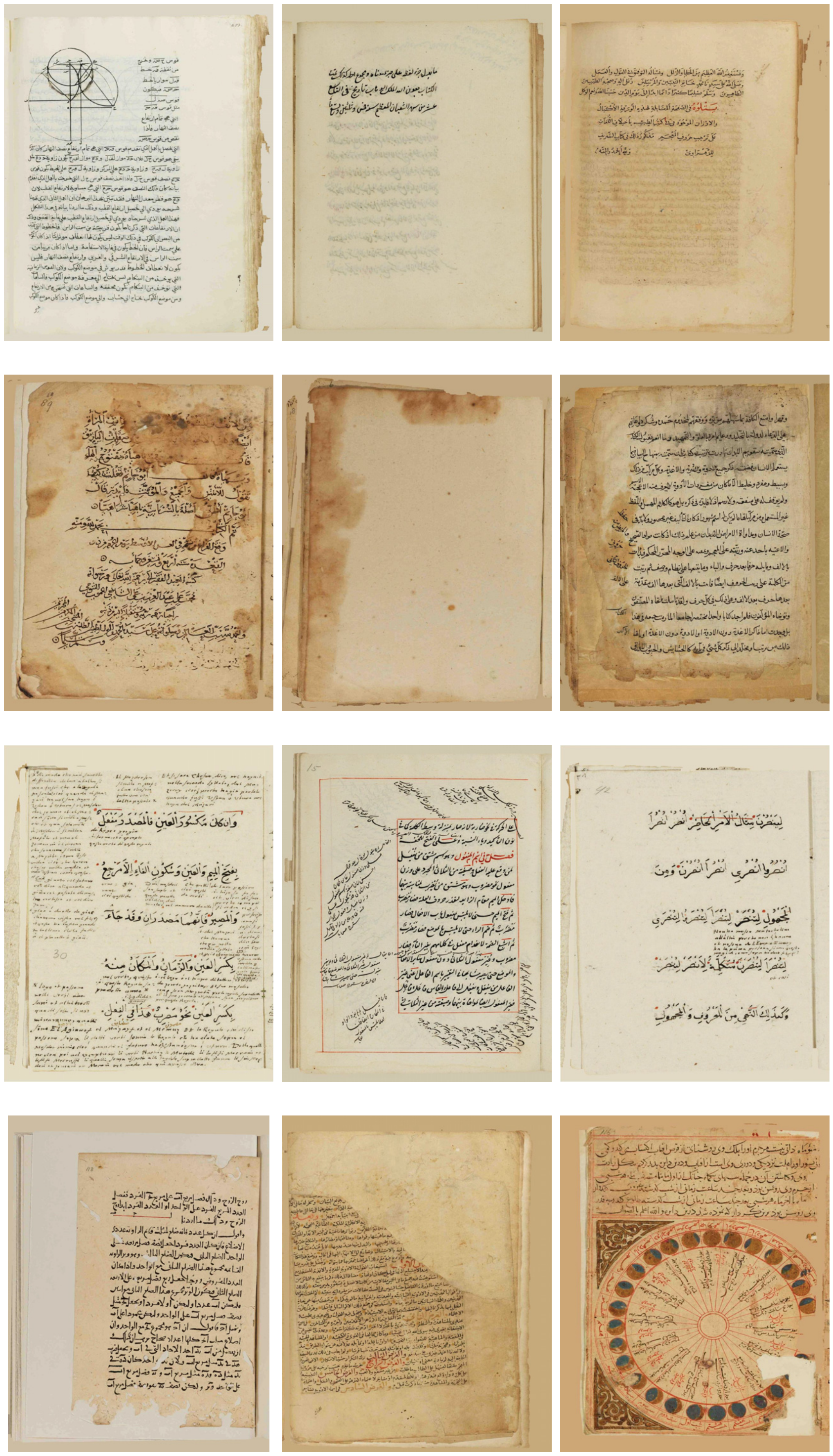


Figure 4.1 Some representative samples of degraded document images selected from different degradation categories in the MHDID dataset: paper translucency, stains, readers' annotations and worn holes. These degradations are shown from the first row to the forth row, respectively.

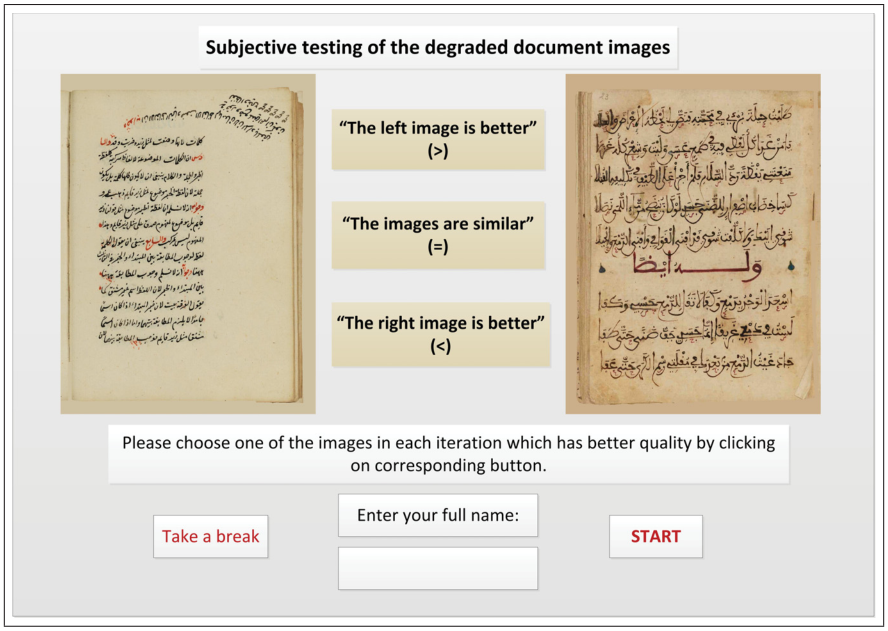
- Ranking: a set of degraded documents, obtained from the same source content, are displayed to subjects who are asked to rank them.

- Pair comparison rating (PCR): it is has been introduced to overcome some limitations of the standard subjective IQA approaches for example the fact that an arbitrary score scale has to be defined. This may lead to ambiguous scores in many cases. PCR offers also the possibility to consider stimuli with similar level of quality. In our experiment we make use of PCR. At each trial, two document images are shown to the subjects. The subjects are asked to compare the two images. A score of 1 is assigned to the image with higher quality and -1 to the other. If the observer could not perceive any difference between the two images, a score of 0 is assigned to both images.

4.2.3 Test environment and data collection

For the purpose of visual evaluation of the document images, 18 graduate and 18 undergraduate students were asked to judge the quality of degraded documents. The subjects gave their rating according to the overall quality of the document images. They were all given basic information on what *physical noise* means in document images. The subjects were asked to ignore other common properties of the document images such as font size, font type, text direction, density of text/non-text, etc. Those degradations that are closer to the text regions are often more annoying because of the reduced readability. PCR method was used for comparison of document images instead of the OCR performance (Sun *et al.*, 2017). In each comparison, subjects may choose one of the three options, “The left image is better", “The images are similar" and “The right image is better". Fig. 4.2 shows the graphical user interface used in the experiment.

This experiment was done in an unconstrained and normal conditions. Identical desktops were used for subjective tests. Each of them had 16 GB RAM and 64-bit Windows operating system. These desktops were placed in a laboratory with normal indoor light with calibrated 17-inch LCD monitors. Table 4.1 indicates the detailed information about the previously proposed dataset (VDIQA) (Shahkolaei *et al.*, November 2017), and the MHDID dataset.


Subjective testing of the degraded document images

Please choose one of the images in each iteration which has better quality by clicking on corresponding button.

Figure 4.2 Screenshot of the interface used for subjective evaluation.

Table 4.1 Comparison between VDIQA and MHDID datasets. MTD stands for Multiple Types of Distortions.

Database	VDIQA	MHDID
Type of database	MTD	MTD
Ref. images available	No	No
Data format	MOS	MOS
Number of images	177	335
Subjective evaluation	PCR	PCR
Size of images	1400 × 2000	1024 × 1280
Range of scores	[0 9]	[0 9]
Number of subjects	27	30
Image format	JPG	JPG
Number of ratings for each image	10	15
Separation of degradation types	No	Yes
Total size of dataset	∼58MB	∼53MB
Screen	17"LCD	17"LCD

Each degraded document with a border degradation was compared with other degraded documents at least 15 times. We removed the judgments of six subjects as being outliers. These subjects had maximum disagreement rate with other subjects. Then, the judgments given by 30 observers are normalized between 0 and 9 as follows:

$$z_i = 9 \times \frac{x_i - min(x)}{max(x) - min(x)} \tag{4.1}$$

where $x = (x_1, ..., x_n)$ and z_i is the i^{th} normalized data. A quality score of 0 corresponds to the lowest perceptual quality and 9 corresponds to the highest perceptual quality.

As mentioned before, each pair is compared by at least 15 subjects. For i^{th} distorted image, its MOS value is calculated by:

$$\text{MOS}_i = \frac{\sum_{j=1}^{N_i} C_{i,j}}{N_i} \tag{4.2}$$

where N_i denotes the number of pair comparisons with i^{th} image involved, and $C_{i,j} \in \{-1, 0, 1\}$ is the outcome of the j^{th} pair comparison for the i^{th} image.

The histogram of MOS values in the MHDID dataset is shown in Fig. 4.3. It is clear that the MOS values are somewhat evenly distributed for highly degraded, moderately degraded, and lightly degraded document images in the built dataset.

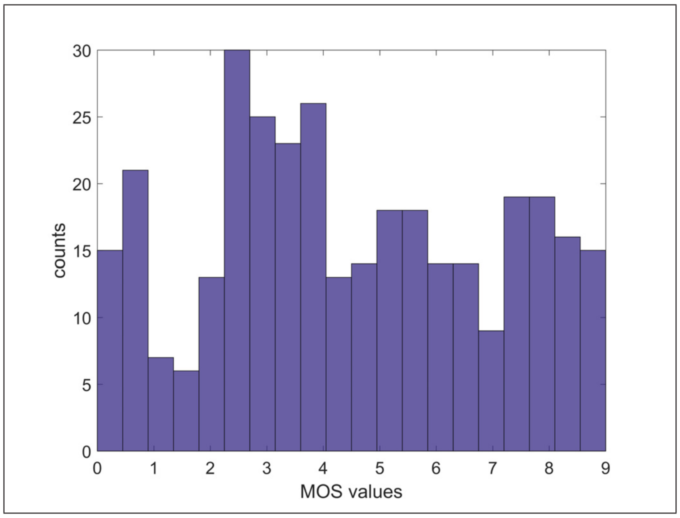


Figure 4.3 Histogram of MOS values in the developed dataset.

4.3 Dataset analysis

In order to characterize the source images in our dataset, we compute two descriptors related to the colorfulness attribute and the edginess contained in the images. Other image low-level features (Lahouhou *et al.*, 2010) could be used to characterize the dataset.

The colorfulness (CF) indicates the variety and intensity of colors in the image. For an RGB color image, it is calculated by using the following formula (Hasler & Susstrunk, 2003):

$$\mathrm{CF} = \sqrt{\sigma_{rg}^2 + \sigma_{by}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{by}^2}, \tag{4.3}$$

where $rg = R - G$ and $by = 0.5(R+G) - B$. σ_k and μ_k for $k \in [rg, by]$ stand for the standard deviation and mean of the opponent color components rg and by, respectively.

The spatial information (SI) is related to the edginess. We use the expression provided in (Winkler, 2012) and adapted to the size of our dataset images.

$$\mathrm{SI} = \sqrt{\frac{L}{1280}}\sqrt{\sum_{i=1}^{N}\frac{(s_h^2(i) + s_v^2(i))}{N}}, \tag{4.4}$$

where s_h and s_v denote the horizontal and vertical Sobel filtered luminance components, respectively. L and N are the numbers of lines and the number of pixels in the filtered image. It should be mentioned that the value of 1280 is the maximum size (in vertical direction) of the considered images in the dataset.

The scatter plot of SI versus CF for 335 images in MHDID dataset is exhibited in Fig. 4.4. This plot shows that while document images are selected from the same library, how much their content differs in terms SI and CF. Note that there is no direct relation with the SI and the document images degradations. SI is an index of the edge energy for the whole image regardless of being true edges of text or degradation edges. It could be noticed that the MHDID dataset exhibits a moderate to low dynamic range of color and edginess compared to natural image database as those analyzed in (Winkler, 2012; Qureshi *et al.*, 2017). Moreover, the overall scatter plot of SI and CF is somewhat balanced. This nice property reveals richness and diversity of the image database.

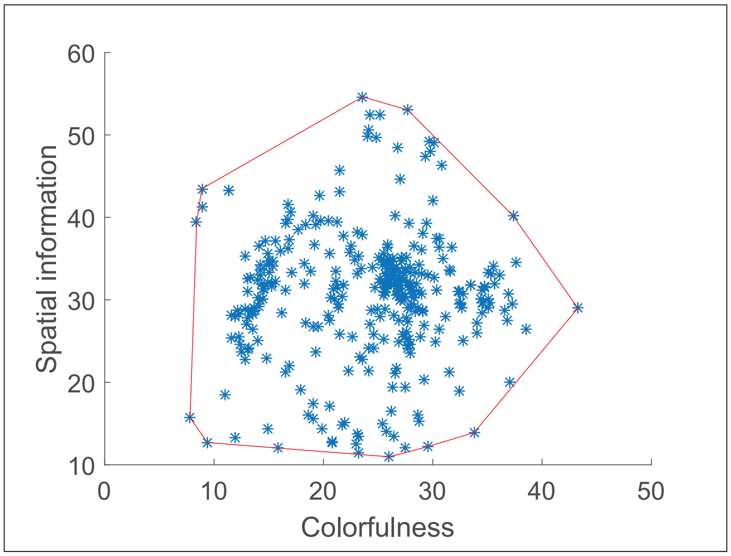


Figure 4.4 A scatter plot of spatial information (SI) against colorfulness (CF) for 335 degraded document images in the MHDID dataset.

4.4 Conclusion

In this paper, a new database has been developed for the purpose of visual quality evaluation of historical document images. The dataset, called MHDID, consists of 335 degraded document images with four types of physical degradations. The document images in MHDID have diverse image content in terms of the edge and color information. These two properties of the developed dataset provide a challenging benchmark for a fair evaluation of the document image quality assessment metrics. In future work, we will include more images as well as more types of physical degradations in the dataset. This dataset could be useful to pave the way for further research on the automatic quality assessment, design of new IQA metrics, degradation modeling and classification of damaged manuscripts.

Acknowledgment

This publication was made possible by NPRP grant NPRP 7-442-1-082 from the Qatar National Research Fund (a member of Qatar Foundation). The statements are solely the responsibility of the authors.

CHAPTER 5

ARTICLE III- BLIND QUALITY ASSESSMENT METRIC AND DEGRADATION CLASSIFICATION FOR DEGRADED DOCUMENT IMAGES

Atena Shahkolaei[1], Azeddine Beghdadi[2], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Department of Computer Science and Engineering, Paris 13 University, Paris, France

Published in Signal Processing: Image Communication
Volume 76, August 2019, Pages 11-21

Abstract

Epochal documents suffer from several types of noises that accumulate and evolve over time. This significantly affects their quality and makes their storage and the interpretation of their visual content problematic.The digital preservation seems the most viable and the most promising. Moreover, measuring the amount of degradation and quality assessment of degraded documents is highly desirable for applications such as selecting the proper algorithms for enhancement and analysis of document images, filtering the damaged images, tuning the processing algorithms parameters, document repairing, psychological study, etc. The first contribution of this work is the proposition of an efficient Multi-distortion Document Quality Measure (MDQM) for quality assessment of physically degraded document images. The proposed MDQM metric is based on three sets of spatial and frequency image features. These features are extracted from two layers of text and non-text and mapped to the mean opinion (MOS) values using regression function. The second contribution of this work is to estimate the probability of four common distortion types, namely, paper translucency, stain, readers annotations and worn holes in the degraded images. In our experiment, the correlations of seven no-reference image quality assessment (NR-IQA) metrics with the MOS values are evaluated on two available datasets. It is shown that the performance of MDQM metric is significantly better than the state-of-the-art NR-IQA metrics. Moreover, the experimental results demonstrate that MDQM metric not only leads to high efficacy for classification of the various degradations but also maintains a remarkable run-time efficiency.

Keywords
No-reference image quality assessment, degraded document images, physical noises, local phase, degradation classification, support vector machine.

5.1 Introduction

Historical document images constitute an important country's cultural heritage and civilization. Therefore, maintaining these cultural heritages is of great importance, and it is the responsibility of each government. In recent decades, digitizing historical documents and manuscripts to preserve and make them accessible via electronic media has received a considerable amount of attention (Eilertsen *et al.*, 2013). Although the issue of digitizing these documents is mostly solved, the problem of analyzing them is still an ongoing challenge.

The image quality of these documents can be assessed subjectively and objectively. Although subjective assessment is the most accurate assessment technique, it demands human participants which makes the assessment time-consuming, tedious and expensive. Therefore, objective assessment is the primary choice of image quality assessment (IQA) applications. Objective IQA automates the estimation of image quality by substituting the human perception process with some quality metrics. Objective assessment methods can be classified into three main categories according to the availability of the reference image. These categories are (1) full reference (FR), (2) reduced-reference (RR) and (3) no reference (NR) (Zhang *et al.*, 2011). For FR metrics both original and distorted images are available (Nafchi *et al.*, 2015; Sheikh & Bovik, 2006; Wang *et al.*, 2004). RR metrics use the partial information about both the reference and degraded images (Gu *et al.*, 2013; Rehman & Wang, 2012; Li & Wang, 2009). Finally, for the NR methods, the evaluation of quality is based on some features and properties of the degraded image without referring to the original one. However, very often a priori knowledge of the distortions is used in the design of NR-IQA (Moorthy & Bovik, 2011; Saad *et al.*, 2012; Moorthy & Bovik, 2010; Gu *et al.*, 2015; Mittal *et al.*, 2012).

In recent years, several objective NR-IQA metrics have been proposed in the literature for different applications. In the following, we provide a brief review of NR-IQA metrics.

Moorthy et al. (Moorthy & Bovik, 2011) proposed Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) for NR-IQA. In their framework, first, scene statistics of the natural images are extracted to classify the images into different distortion types. Then, the extracted statistics are used to assess the distortion-specific quality. The final quality score is found by a combination of the results from the aforementioned two steps. The BLind Image Integrity Notator using DCT statistics (BLIINDS-II) (Saad *et al.*, 2012) metric is based on finding the statistics of fitting a Generalized Gaussian Distribution (GGD) model to the Discrete Cosine Transform (DCT) coefficients. A generalized Natural Scene Statistics (NSS) was used for local DCT coefficients. Then, it transforms the model parameters into features. The popular BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) (Mittal *et al.*, 2012) metric takes advantage of the natural images statistics in the spatial domain. This metric computes the statistics of fitting the Mean Subtracted Contrast Normalized (MSCN) coefficients by symmetric and asymmetric GGDs in two scales. No transform (e.g., DCT, wavelet, etc) were not used in this metric. The Curvelet Quality Assessment (CurveletQA) (Liu *et al.*, 2014) metric finds the statistical correlations between the curvelet energy distributions and distortions of the image using a log-pdf curvelet model. The High Order image Statistics (HOS-DIQA) (Xu *et al.*, 2016) metric is proposed for Document Image Quality Assessment (DIQA) applications. It extracts a codebook from local image patches by using the k-means clustering. Then, the differences between the statistics of local features and these codewords are calculated as the quality metric. The Blind Quality Measure for SCIs (BQMS) (Gu *et al.*, 2016) is the first proposed metric for quality assessment of the screen content images that exploit perceptual features according to the free energy measure and structural degradation information. In this metric, an immense number of training data was utilized in order to avoid the overfitting in comparison with other IQA metrics. NR Free Energy-Based Robust Metric (NFERM) is an

NR-IQA metric which was proposed based on the free energy principle and important HVS inspired features followed by an SVM-based regression module (Gu *et al.*, 2015). Indeed, the extracted features can be classified into three groups in the NFERM metric: the features inspired by the free energy principle and the structural degradation model, HVS-inspired features and the possible losses of naturalness in the distorted image.

It is worth noticing that, the majority of the objective image quality assessment metrics work on the entire image information for evaluating the quality of images (Liu *et al.*, 2014; Saad *et al.*, 2012; Moorthy & Bovik, 2011; Gu *et al.*, 2016; Ghadiyaram & Bovik, 2017; Gu *et al.*, 2015). Therefore, the use of these metrics in the case of document images does not seem adequate. Indeed, the observer's visual attention is not taken into account in the sense that the effect of visual attraction by the text is not exploited. In recent years, some objective IQA measures based on different layers of images have been proposed for predicting the quality of images. Shahkolaei et al. (Shahkolaei *et al.*, November 2017) proposed the VDQAM metric for quality assessment of degraded document images. This measure is based on the statistical analysis of a set of selected features extracted from four layers of the image using a log-Gabor filters based image segmentation method. In the same vein, another approach based on a background/foreground image segmentation and a patch selection scheme has been proposed recently for the design of a new IQA (Alaei *et al.*, 2017).

In order to maintain, control and enhance the quality of degraded documents and also decrease the negative effect of distortions on diverse processing and analysis systems, it is necessary to estimate the probability of different distortion types in ancient document images. In the literature, degradation classification methods can be considered as a variational or statistical approaches (Moghaddam & Cheriet, 2010). To classify the degradation types in natural images, Moorthy and Bovik (Moorthy & Bovik, 2010) introduced a two-step framework for blind image quality evaluation based on the natural scene statistics (NSS). This metric estimates the probability of different degradation types in natural images such as white noise, Gaussian blur, JPEG, etc. In (Chetouani *et al.*, 2012), a degradation classification method for natural images was proposed that is based on the recognition accuracy of degradation type and overall image quality assessment. A Bayesian approach was utilized in order to predict the type of distortion in images using image quality metrics in (A. Chetouani, 2010, Pages: 714-717). In (Chetouani & Beghdadi, 2011), a global FR image quality measure (IQM) was proposed based on linear discriminant analysis (LDA) classifier and neural approach. In this method, an artificial neural network (ANN) was used to improve the performances of the existing IQMs by selecting and combining the best IQMs for each considered degradation. Recently, Xiongkuo Min et al. (Min *et al.*, 2018) proposed a blind pseudo reference image (BPRI) metric in order to estimate some distortions such as blockiness, sharpness and noisiness. This method is training-free expect for degradation classification step.

Recently, two degraded document datasets with the human opinion scores were introduced in the literature (Shahkolaei *et al.*, November 2017; A. Shahkolaei, A. Beghdadi, S. Al-maadeed and M. Cheriet, 2018). In these datasets, pair comparison rating was used instead of the previously used OCR performance (Sun *et al.*, 2017). Visual document image quality assessment (VDIQA) dataset is the first dataset of degraded document images along with human opinion scores for each document image (Shahkolaei *et al.*, November 2017). This dataset contains 177 historical document images with various degradations. It should be mentioned that the number of degraded documents in this dataset is not adequate for some future goals such as degradation classification and modeling. Also, no information about the type of degradations is provided in this dataset. Therefore, the MHDID dataset was introduced by Shahkolaei et al. (A. Shahkolaei, A. Beghdadi, S. Al-maadeed and M. Cheriet, 2018) to overcome these limitations. This dataset consists of 335 degraded documents with four types of degradation categories, namely, paper translucency, stain, readers annotations and worn holes. With the availability of such dataset, more researches on degradation classification and modeling can be conducted in addition to the document image quality assessment.

In this paper, we propose a blind image quality assessment metric (MDQM) for quality assessment of damaged manuscripts. The proposed metric assumes that the sensitivity of the human visual system varies depending on the observed region in the document image (background and foreground). The proposed IQA measure not only uses the existing statistical features, but it introduces a new feature, namely the local phase that follows the Gaussian distribution. To the best of our knowledge, for the first time, a degradation classification algorithm is also proposed to estimate the probability of four common types of degradation in ancient document images.

This paper is organized as follows. Section 6.3 provides the detailed description of the proposed IQA measure. In section 5.3, we outline the degradation classification model. Section 6.4 provides experimental results followed by a conclusion in section 7.3.

5.2 Proposed Document Image Quality Measure (DIQM)

The proposed method is based on some observations and specificities of the document images related to the way the human perceives structured images. In the literature, many researchers have studied the statistics of natural images and their correlation with the Human Visual System (HVS) (Geisler, 2008; Mittal *et al.*, 2012; Zhang & Chandler, 2013; Ghadiyaram & Bovik, 2017). One major difference between a document image and a natural image is the existence of foreground and background that can be easily distinguished by the human visual system. Indeed, the document image is considered as a two-phase media where the object/target are identified and recognized instantaneously. We can mention the various types of noises that exist in the natural and historical document images as the second difference. Previous study (Shahkolaei *et al.*, November 2017) shows that document IQA on a properly segmented document image is a good strategy to design better performing metrics. A novel NR-IQA metric for quality assessment of old document images based on this approach is proposed.

Fig. 5.1 illustrates the framework of the proposed blind image quality measure (MDQM). The features used in MDQM metric can be grouped into three major classes: *i*) the first corresponds to the mean of the local phase of the three color image channels, *ii*) the second group of features is composed of the MSCN coefficients and MSCN coefficients of gradient information from two layers of foreground and background, and *iii*) statistical features such as mean and standard deviation are extracted from the MSCN coefficients and gradient of MSCN coefficients in the third group. Support vector regression (SVR) is utilized to map the extracted features to MOS values. In the following the main steps of the proposed method are described and discussed.

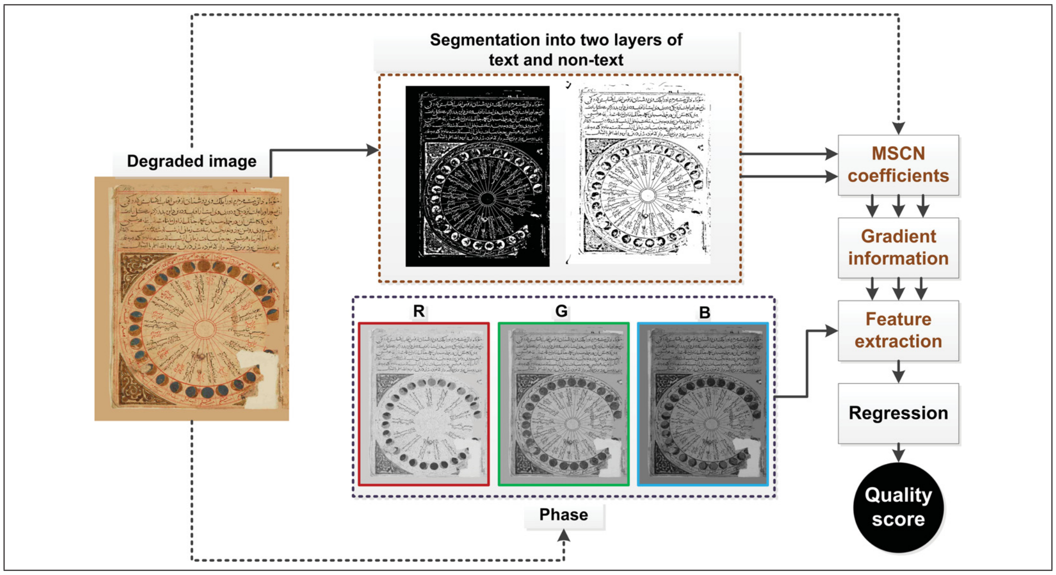


Figure 5.1 The framework of the proposed blind MDQM metric.

5.2.1 Text/non-text image segmentation

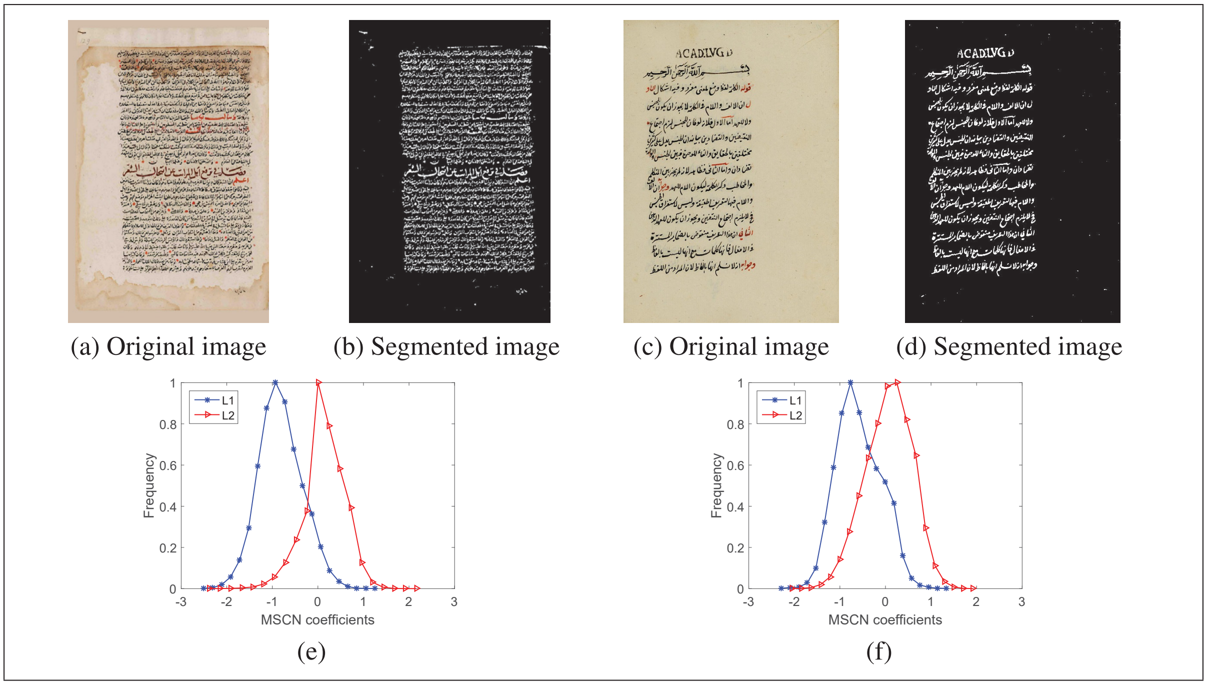
It is worth to mention that the human visual system responses to the degradations appearing in the foreground and background of a given image are quite different. In other words, due to various types of noise and other distortions in text and non-text regions, the sensitively of human vision to different parts of document images is not the same. Therefore, it is of great importance to segment document images into different layers based on the HVS sensitivity, because human perceptual observation is the ultimate judge of the image quality evaluation process.

Recently, a new blind IQA metric dedicated to document images and based on image segmentation has been proposed in (Shahkolaei *et al.*, November 2017). The observed degraded document image is segmented into four layers namely, text, degradations far from the text, degradations close to the text, and non-degraded pixels. Segmenting the image into four layers by using this approach is not always effective, because there is no guarantee that each layer is non-empty. Therefore, the feature analysis of these layers based on their MSCN coefficients histograms may not be accurate. To overcome this limitation, the segmentation of the degraded document image is restricted to two layers, namely text and non-text zones. The segmentation method is described below.

Simple global thresholding and other spatial domain filters cannot be used for the segmentation of degraded documents into different layers, due to the local variations in degraded document images (Shahkolaei *et al.*, November 2017). The local adaptive methods can be used for segmentation of degraded document images. Log-Gabor filters have been reported to show good performance in many applications ranging from image quality assessment to image segmentation (Bourgeois *et al.*, 2004; Allier & Emptoz, 2003; Jain & Farrokhnia, 1990; Shahkolaei *et al.*, November 2017; Zhang & Chandler, 2013).

Different pairs of the symmetric and anti-symmetric functions, such as the Gabor, Log-Gabor, Gaussian derivatives, the difference of Gaussians, and Cauchy functions, which are known as quadrature pair filters were used in the literature (Papari & Petkov, 2011; Boukerroui *et al.*, 2004). In this work, we exploit the joint spatial and frequency localization of log$-$Gabor filter for segmenting the historical document images into two layers, namely foreground and background. Let quadratic pairs are defined by $M^e_{\rho r}$ and $M^o_{\rho r}$ which define the even-symmetric and odd-symmetric wavelets at a scale ρ and orientation r (Papari & Petkov, 2011). By considering $I(\mathbf{x})$ as a two-dimensional signal, the response associated with each quadratic pair of filters at each image point $\mathbf{x}$ forms a response vector obtained by convolving the wavelet filters with $I(\mathbf{x})$ (Shahkolaei *et al.*, November 2017):

$$[e_{\rho r}(x), o_{\rho r}(x)] = [I(\mathbf{x}) * M^e_{\rho r}, I(\mathbf{x}) * M^o_{\rho r}] \tag{5.1}$$



(a) Original image (b) Segmented image (c) Original image (d) Segmented image

(e) (f)

Figure 5.2 Illustrations of the proposed document image segmentation method and the associated MCSN coefficients histograms of text and non-text classes. Foreground and background pixels of two degraded images are shown in (b) and (d) images with white and black color, respectively. In (e) and (f) MCSN histograms where the blue plot is related to the text regions (L_1), while the red plot is related to the non-text regions (L_2).

where * denotes convolution, and $e_{\rho r}(x)$ and $o_{\rho r}(x)$ are the real and imaginary responses in the complex-valued frequency domain. The signal energy is defined as the summation of the responses across the five scales and directions as follows:

$$\mathrm{E(x)} = \sum_{\rho r} e_{\rho r}(x) \tag{5.2}$$

The degraded document image segmented into two layers, namely text and non text pixels, using a thresholding process defined as follows: (Shahkolaei *et al.*, November 2017):

$$\begin{cases} L_1: \ \mathrm{E(x)} \leq \text{mean negative values} \\ L_2: \ \mathrm{E(x)} > \text{mean negative values} \end{cases} \tag{5.3}$$

where, layer L_1 approximates the text regions, while the layer L_2 estimates the non-text pixels.

Two typical degraded document images are shown in Fig. 5.2. It is clear that the proposed segmentation method has remarkable efficacy in separating text and non-text pixels. From Fig. 5.2, it can also be observed that the histograms of MSCN coefficients for both layers of text and non-text of the degraded images exhibit a Gaussian distribution shape.

After segmenting document images into two layers, the statistics of MSCN coefficients and their gradient information are computed for each layer and the entire document image. In the next subsection, more details about these statistics are given.

5.2.2 Document statistical analysis based on the MSCN coefficients

In order to remove the dependency between different image features, MSCN coefficients are used as a normalization technique in the proposed approach. The main behavior of the MSCN coefficients is their tendency toward a Gaussian distribution. MSCN coefficients are commonly used for NR-IQA of natural images (Mittal *et al.*, 2012; Zhang & Chandler, 2013). More details on the relevancy of the MSCN coefficients in the design of the proposed IQA metric are given in the following.

In (Daniel L Ruderman, 1994), MSCN coefficients are computed by applying a local nonlinear operation to a luminance channel of images, to remove local mean displacements of the luminance and normalize the local variances of luminance. Given an image $I(i,j)$, MSCN coefficients can be computed as follows:

$$\mathrm{MSCN}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{5.4}$$

where $i \in 1,2,...,M,\ j \in 1,2,...,N$ are spatial indices; M, N are the image height and width respectively; $C = 1$ is a constant that prevents the denominator to be zero (Zhang & Chandler, 2013). The local mean $\mu(i,j)$ and standard deviation $\sigma(i,j)$ are defined as follows:

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} I_{k,l}(i,j) \tag{5.5}$$

and

$$\sigma^2(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} \left(I_{k,l}(i,j) - \mu(i,j)\right)^2 \tag{5.6}$$

where, $\omega = \{\omega_{k,l} | k = -K,...,K, l = -L,...,L\}$ is a two-dimensional symmetric Gaussian weighting window.

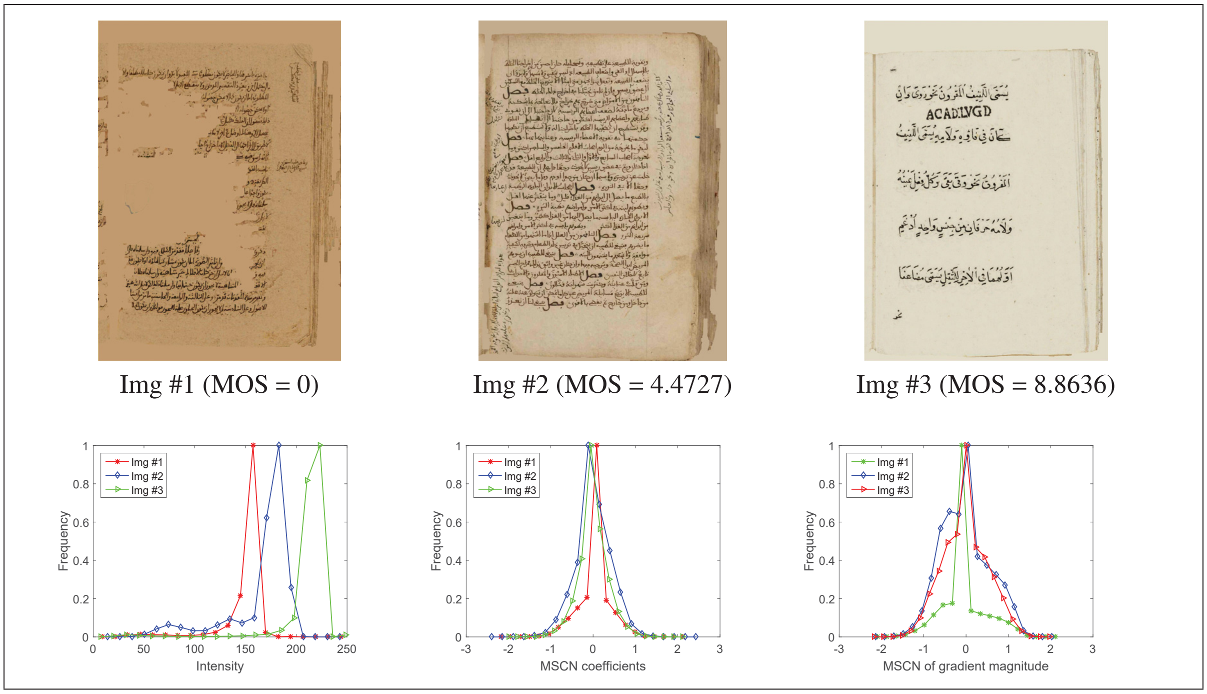
Image gradients contain momentous information about the distribution of edges and variations of local contrast that were utilized extensively in many FR and NR image quality assessment metrics (Nafchi *et al.*, 2015; Liu *et al.*, 2012; Xue *et al.*, 2014; Kundu *et al.*, 2017). In this work, we use the gradient magnitude as explained in the following.

The estimated gradient magnitude M of the image $I(i,j)$ is given by:

$$M(i,j) = \sqrt{(I * H_x)^2(i,j) + (I * H_y)^2(i,j)} \tag{5.7}$$

where $*$ denotes the convolution operator. H_x and H_y are the convolution masks associated with the vertical and horizontal Sobel operator.

In this study, the statistics of MSCN coefficients of gradient magnitude are computed from each layer and the entire degraded image separately. Fig. 5.3 shows the histograms of the intensity, MSCN distributions and MSCN of gradient magnitude for three historical document images with different MOS values from the MHDID dataset. The Gaussian behavior of the MSCN coefficients and its gradient, unlike what is shown in the intensity histogram, can be observed from the plots. Besides, it can be seen that although the histogram of MSCN coefficients and


Img #1 (MOS = 0)

Img #2 (MOS = 4.4727)

Img #3 (MOS = 8.8636)

Figure 5.3 Histograms of intensity, MSCN and MSCN of gradient magnitude for three degraded document images with different MOS values, selected from the MHDID dataset.

its gradient follow the Gaussian distribution, the histogram of MSCN tends more to a bell curve in comparison with the histogram of MSCN of gradient magnitude.

5.2.3 Local phase information of RGB channels

It has been observed that the human visual system reacts more strongly to the points where there is highly ordered phase information (Liu & Laganiare, 2007). Since the human visual cortex is sensitive to phase congruent structures (Henriksson *et al.*, 2009), the phase congruency (PC) value at a location can reflect how likely it is a relevant perceptual salient feature (Zhang *et al.*, 2011). Also, phase carries more visual information than does magnitude (Oppenheim & Lim, 1981). Taking these observations into account, PC was used as a feature in several IQA metrics (Zhang *et al.*, 2011; Saha & Wu, 2013; Krzic *et al.*, 2016; Wang & Simoncelli, 2004; Liu & Laganiare, 2007).

Our proposed IQA metric utilizes features derived from a locally weighted mean phase angle (LWMPA) (Kovesi, 1999), which is robust to noise, through the three color channels (R, G, B). In the following, the basic notions on LWMPA and its use for IQA are briefly described.

Using the equation (5.1), we find the values of $e_{\rho r}(x)$ and $o_{\rho r}(x)$ which leads to find the local phase $\omega_{\rho r}(x)$ at the given scale and orientation of the wavelet. This value is computed by the following formula:

$$\omega_{\rho r}(x) = arctan2(o_{\rho r}(x), e_{\rho r}(x)) \tag{5.8}$$

$$arctan2(i,j) = 2arctan(\frac{i}{\sqrt{i^2+j^2}+j}) \tag{5.9}$$

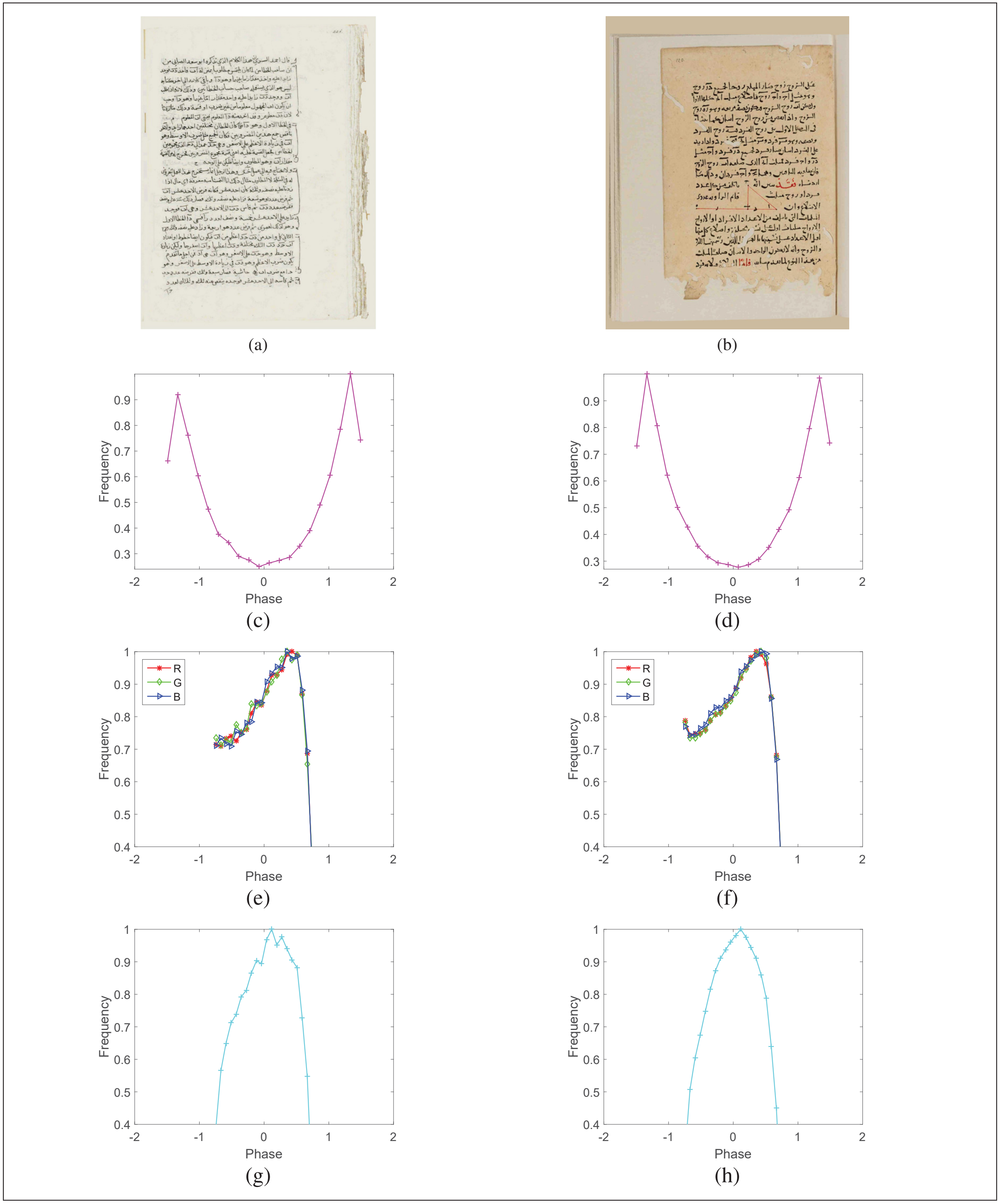
Finally, the LWMPA $ph(x)$ is computed by summing all the response vectors in all the possible orientations and scales:

$$ph(x) = arctan2\left[\sum_{\rho,r} e_{\rho r}(x), \sum_{\rho,r} o_{\rho r}(x)\right] \tag{5.10}$$

LWMPA takes it values within the range [$-\pi/2$, $\pi/2$].

It should be mentioned that for feature extraction, the RGB image components are quite efficient in comparison with the transformed color spaces. The interesting point of the proposed metric is that LWMPA is computed directly from each channel of ancient document images (R, G, B) separately. RGB image composites are quite efficient for feature extraction. Then, the final LWMPA (ph_{RGB}) is calculated from the mean of three LWMPAs for each degraded image (x) and expressed as:

$$ph_{RGB}(x) = \frac{1}{3} \sum_{c=R,G,B} ph_c(x) \tag{5.11}$$



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5.4 Histograms of the local phase of two degraded document images ((a)-(b)). The histograms in the second row indicate the distribution of local phase which is computed directly from gray scale version of images ((c)-(d)). The histograms in the third row correspond to the distribution of local phase on the RGB channels separately ((e)-(f)). The histograms of the mean of local phase of R, G and B channels are shown in by (g) and (h).

Fig. 5.4 illustrates the local phase distribution of two degraded document images in three cases: when the local phase is computed from *i*) gray-scale image, *ii*) individual RGB channels and *iii*) the combination of RGB channels $(ph_{RGB}(x))$ that are proposed in this work. From the plots of Fig. 5.4, it can be seen that the local phase distribution of *iii* fairly tends towards a Gaussian distribution, while *i* and *ii* histograms are non-Gaussian. To the best of our knowledge, it is for the first time that the local phase as a relevant feature for image distortion analysis.

5.2.4 Symmetric and asymmetric generalized Gaussian distribution

The symmetric and asymmetric behaviors of the MSCN, its gradient, and local phase suggest fitting the model into asymmetric and symmetric Gaussian distributions. Extracted features are parametrized using a generalized Gaussian distribution. The zero mean generalized Gaussian distribution (GGD) can be expressed as:

$$f(x;\varphi,\sigma^2) = \frac{\varphi}{2\beta\Gamma(1/\varphi)} e^{\left(-\frac{|x|}{\beta}\right)^{\varphi}} \tag{5.12}$$

where φ controls the shape of the distribution and σ^2 is the variance of the distribution. The parameter β and gamma function Γ are defined as:

$$\beta = \sigma\sqrt{\frac{\Gamma(1/\varphi)}{\Gamma(3/\varphi)}} \tag{5.13}$$

$$\Gamma(\varphi) = \int_0^{\infty} t^{\varphi-1} e^{-t} dt \ , \qquad \varphi > 0 \tag{5.14}$$

When the shape of the histogram is not symmetric, the GGD cannot be used for fitting to that histogram (Sharifi & Leon-Garcia, 1995). Therefore, in this situation, asymmetric generalized Gaussian distribution (AGGD) is more adequate (Mittal *et al.*, 2012). The AGGD with zero-mode is defined by:

$$f(x;\digamma) = \begin{cases} \frac{\vartheta}{(\beta_l+\beta_r)\Gamma(1/\vartheta)}\exp\Big(-(\frac{-x}{\beta_l})^{\vartheta}\Big), & x < 0 \\ \frac{\vartheta}{(\beta_l+\beta_r)\Gamma(1/\vartheta)}\exp\Big(-(\frac{x}{\beta_r})^{\vartheta}\Big), & x \geq 0 \end{cases} \tag{5.15}$$

The parameters β_l and β_r are defined as follows:

$$\beta_l = \sigma_l\sqrt{\Gamma(\frac{1}{\vartheta})/\Gamma(\frac{3}{\vartheta})} \tag{5.16}$$

$$\beta_r = \sigma_r\sqrt{\Gamma(\frac{1}{\vartheta})/\Gamma(\frac{3}{\vartheta})} \tag{5.17}$$

$\digamma$ is the set of three statistical features which can be defined as the following:

$$\digamma = (\sigma_l^2, \sigma_r^2, \vartheta) \tag{5.18}$$

where σ_l^2 and σ_r^2 are the variance of the left side and right side of the distribution, respectively. The parameter ϑ controls the shape of the distribution.

These classical statistical functions provide useful information about the statistical behavior of the distributions. Therefore, two parameters of GGD (φ, σ^2) and three parameters of AGGD ($\sigma_l^2, \sigma_r^2, \vartheta$) on two layers of the degraded image constitute the first part of the feature vector of MSCN coefficients and MSCN of gradient magnitude for the proposed metric. Additionally, two parameters of GGD (φ, σ^2) and three parameters of AGGD ($\sigma_l^2, \sigma_r^2, \vartheta$) on the entire degraded image constitute the second part of the feature vector of MSCN coefficients, MSCN of gradient magnitude and local phase for the MDQM metric.

All the above statistical functions are calculated from two image scales; the original image scale and its reduced resolution by the "bicubic" function. The final feature vector has 124 features (62 at each scale) for each image.

5.3 Degradation identification of degraded document images

Identification of the type of distortion and estimation the amount of each degradation can be specified where applicable to be used for degradation modeling. In addition, it can be useful for the optimization of enhancement parameters. For instance, parameters of the selected binarization methods can be properly adjusted by the provided information about the possible degradation types. In this section, the estimation of the probability of each distortion is provided for degraded document images in MHDID dataset.

In order to estimate the probabilities, we need to feed the previously extracted features to Support vector machine (SVM) classifier. SVM classifier is more popular in comparison with other types of classification techniques because it performs well in high-dimensional spaces, avoid over-fitting and have good generalization capabilities (Vapnik, 2000). Therefore, a multi-class SVM classifier is used to classify the degraded images into four different distortion categories _ paper translucency, readers' annotations, stain and worn holes in this work.

For the classification stage, the images of each distortion category are subdivided into 80% for training and 20% for testing sets. The degradation types are denoted as κ_i, $i = \{1,...,4\}$, where each number corresponds to a type of distortion. A degree two polynomial is utilized as the kernel function to estimate the probability of each distortion type. The greatest value of the probabilities indicates the dominant distortion type of a historical document image.

For SVM classification, LIBSVM toolbox (Chang & Lin, 2011) is used. Classification probabilities for four degraded document images on the newly proposed quality metric for degraded documents (VDQAM) (Shahkolaei *et al.*, November 2017) and our proposed metric (MDQM) are shown in Fig. 5.5. The histograms in green color demonstrate the ability of MDQM metric for probability estimation of each distortion type in MHDID dataset, while red histograms show this probability on VDQAM metric. This figure clearly illustrates the superiority of the MDQM over VDQAM for degradation classification and probability estimation. For instance, MDQM truly estimates a higher probability of paper translucency for the image Fig. 5.5(a) which contains paper translucency. However, class probability estimations of VDQAM are not

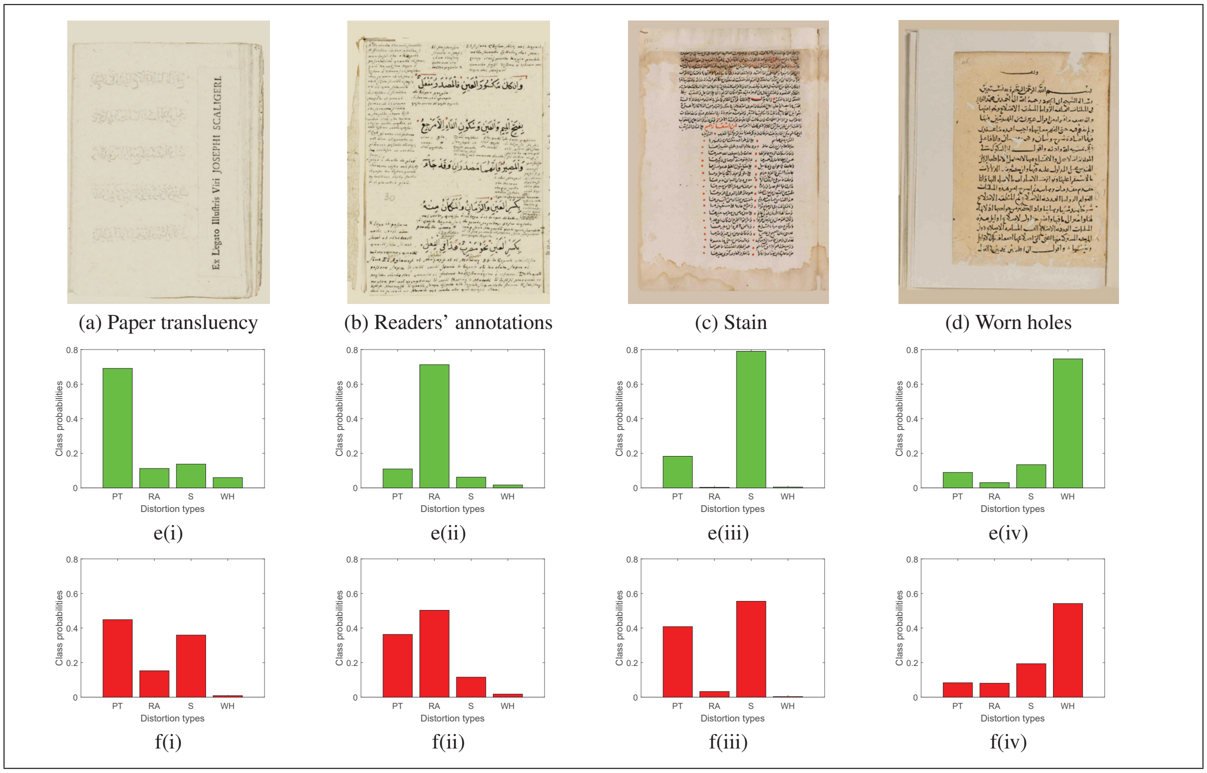

(a) Paper transluency
(b) Readers' annotations
(c) Stain
(d) Worn holes
e(i)
e(ii)
e(iii)
e(iv)
f(i)
f(ii)
f(iii)
f(iv)

Figure 5.5 The performance of degradation classification for four ancient document images by two metrics, VDQAM and MDQM. The first raw (e(i)-e(iv)) and second raw (f(i)-f(iv)) bars show the MDQM and VDQAM operation for the probability estimation of different distortions types, respectively. The X-axis of bars demonstrates different distortion types in the degraded images of MHDID dataset: paper transluency (PT), readers' annotations (RA), stains (S) and worn holes (WH).

accurate for the same image. Moreover, some other degradations such as readers' annotations and stain are remarkably detected by VDQAM metric in Fig. 5.5 f(i), while the amount of these degradations is not significant in the corresponding image (Fig. 5.5(a)).

5.4 Experimental results and discussion

In this section, the performance of the proposed metric is analyzed regarding its ability to predict subjective ratings of image quality on two datasets. To map our feature vectors to MOS values, we fed them into a support vector regression (SVR) process (Basak *et al.*, 2007).

The relationship between objective and subjective quality scores is nonlinear in general. Therefore, a regression is utilized to remove this non-linearity (Sheikh *et al.*, 2006). The reported PCC values in this paper were computed after mapping the quality scores to MOS using the following logistic function:

$$f(x) = \beta_1\left(\frac{1}{2} - \frac{1}{1+e^{\beta_2(x-\beta_3)}}\right) + \beta_4 x + \beta_5 \tag{5.19}$$

where β_1, β_2, β_3, β_4 and β_5 are fitting parameters computed by minimizing the mean square error between quality predictions x and subjective scores MOS.

5.4.1 Performance comparison of the proposed metric

The proposed blind metric (MDQM) is compared with six state-of-the-art metrics on two datasets. The degree of concordance between the subjective and objective evaluations is analyzed in terms of three correlation measures: the Spearman Rank-order Correlation coefficient (SRC), the Pearson linear Correlation Coefficient (PCC), and the Kendall Rank Correlation coefficient (KRC). The PCC and SRC metrics measure prediction linearity and monotonicity, respectively. The KRC is used to evaluate the degree of similarity between quality scores and MOS. A good objective quality metric is expected to attain high values in SRC, PCC and KRC. We evaluated our proposed metric on the VDIQA and MHDID datasets. The results were reported based on the median value of 1000 times train-test for three cases of 20% train 80% test, 50% train 50% test and 80% train 20% test. The median SRC, PCC and KRC values across these 1000 train-test trials are tabulated in Tables 5.1-5.3. The top two metrics are highlighted in the tables. The correlations between objective image quality measures and MOS values show that the MDQM metric outperforms all the other NR-IQA metrics on both datasets.

Table 5.1 Performance comparison of the proposed metric (MDQM) and six blind image quality assessment metrics on the VDIQA and MHDID datasets for 20% train and 80% test. Note the metrics that are specifically designed to assess the physical noises.

Database	Algorithms	Physical noise	20%-80%		
			SRC	PCC	KRC
VDIQA	DIIVINE (Moorthy & Bovik, 2011)	No	0.6755	0.6925	0.4878
	BLIINDS-II (Saad *et al.*, 2012)	No	0.5871	0.6062	0.4235
	BRISQUE (Mittal *et al.*, 2012)	No	0.5648	0.5926	0.4028
	CurveletQA (Liu *et al.*, 2014)	No	0.5052	0.5511	0.3608
	BQMS (Gu *et al.*, 2016)	No	0.6610	0.6823	0.4804
	VDQAM (Shahkolaei *et al.*, November 2017)	Yes	**0.7060**	**0.7204**	**0.5181**
	MDQM	Yes	**0.7224**	**0.7491**	**0.5323**
MHDID	DIVIINE (Moorthy & Bovik, 2011)	No	0.6680	0.7004	0.4860
	BLIINDS-II (Saad *et al.*, 2012)	No	0.6091	0.6357	0.4339
	BRISQUE (Mittal *et al.*, 2012)	No	0.6290	0.6680	0.4494
	CurveletQA (Liu *et al.*, 2014)	No	0.5552	0.6161	0.3931
	BQMS (Gu *et al.*, 2016)	No	0.6302	0.6692	0.4509
	VDQAM (Shahkolaei *et al.*, November 2017)	Yes	**0.6769**	**0.7102**	**0.4864**
	MDQM	Yes	**0.7286**	**0.7637**	**0.5349**

Table 5.2 Performance comparison of the proposed metric (MDQM) and six blind image quality assessment metrics on the VDIQA and MHDID datasets for 50% train and 50% test. Note the metrics that are specifically designed to assess the physical noises.

Database	Algorithms	Physical noise	50%-50%		
			SRC	PCC	KRC
VDIQA	DIIVINE (Moorthy & Bovik, 2011)	No	0.7726	0.7878	0.5795
	BLIINDS-II (Saad *et al.*, 2012)	No	0.6837	0.7008	0.5105
	BRISQUE (Mittal *et al.*, 2012)	No	0.7356	0.7539	0.5466
	CurveletQA (Liu *et al.*, 2014)	No	0.6053	0.6419	0.4437
	BQMS (Gu *et al.*, 2016)	No	0.7573	0.7815	0.5666
	VDQAM (Shahkolaei *et al.*, November 2017)	Yes	**0.7914**	**0.8095**	**0.6095**
	MDQM	Yes	**0.8300**	**0.8429**	**0.6432**
MHDID	DIVIINE (Moorthy & Bovik, 2011)	No	0.7517	0.7838	0.5673
	BLIINDS-II (Saad *et al.*, 2012)	No	0.7256	0.7490	0.5360
	BRISQUE (Mittal *et al.*, 2012)	No	0.7458	0.7783	0.5549
	CurveletQA (Liu *et al.*, 2014)	No	0.6339	0.6930	0.4628
	BQMS (Gu *et al.*, 2016)	No	0.7464	0.7779	0.5548
	VDQAM (Shahkolaei *et al.*, November 2017)	Yes	**0.7686**	**0.7878**	**0.5770**
	MDQM	Yes	**0.8098**	**0.8349**	**0.6219**

5.4.2 Performance comparison of degradation classification model

As mentioned before, each degraded document image belongs to one of the four distortion categories of degradation in the MHDID dataset. For modeling the classification types, SVM classifier was used. Experimental results are done using 10-fold cross-validation setup where all samples were randomly divided into ten folds. In each testing cycle, one fold is used for

Table 5.3 Performance comparison of the proposed metric (MDQM) and six blind image quality assessment metrics on the VDIQA and MHDID datasets for 80% train and 20% test. Note the metrics that are specifically designed to assess the physical noises.

Database	Algorithms	Physical noise	80%-20%		
			SRC	PCC	KRC
VDIQA	DIIVINE (Moorthy & Bovik, 2011)	No	0.7961	0.8292	0.6144
	BLIINDS-II (Saad *et al.*, 2012)	No	0.7218	0.7590	0.5560
	BRISQUE (Mittal *et al.*, 2012)	No	0.7950	0.8283	0.6120
	CurveletQA (Liu *et al.*, 2014)	No	0.6524	0.6985	0.4903
	BQMS (Gu *et al.*, 2016)	No	0.7789	0.8213	0.5931
	VDQAM (Shahkolaei *et al.*, November 2017)	Yes	**0.8176**	**0.8481**	**0.6492**
	MDQM	Yes	**0.8447**	**0.8768**	**0.6785**
MHDID	DIVIINE (Moorthy & Bovik, 2011)	No	0.7862	0.8191	0.6099
	BLIINDS-II (Saad *et al.*, 2012)	No	0.7582	0.7891	0.5723
	BRISQUE (Mittal *et al.*, 2012)	No	0.7894	0.8245	0.6026
	CurveletQA (Liu *et al.*, 2014)	No	0.6623	0.7250	0.4914
	BQMS (Gu *et al.*, 2016)	No	0.7863	0.8266	0.5998
	VDQAM (Shahkolaei *et al.*, November 2017)	Yes	**0.7956**	**0.8294**	**0.6115**
	MDQM	Yes	**0.8395**	**0.8667**	**0.6566**

testing and the rest of them are utilized for the training part. The process is repeated ten times and the final performance is calculated by the average result over all testing cycles.

F-measure values are calculated to evaluate the performance of the proposed metric for degradation detection. The comparison of F-measure values for seven NR-IQA algorithms is shown in Table 5.4. For each degradation type, the top two NR-IQA metrics are highlighted. As observed from the Table 5.4, the proposed metric (MDQM) has the highest performance for detecting different physical noises in the MHDID dataset.

Table 5.4 The comparison of F-measure values for different degradation types in MHDID dataset for the seven NR-IQA metrics. We highlight the top two best performance metrics with boldface. Values are in percent.

F-measure				
Algorithms	Paper translucecy	Redears' annotations	Stains	Worn holes
DIIVINE (Moorthy & Bovik, 2011)	8.70	30.00	56.72	75.00
BLIINDS-II (Saad *et al.*, 2012)	17.39	0	53.13	89.33
BRISQUE (Mittal *et al.*, 2012)	11.76	59.26	57.14	68.29
CurveletQA (Liu *et al.*, 2014)	0	0	57.58	77.78
BQMS (Gu *et al.*, 2016)	8.33	10.53	48.28	42.42
VDQAM (Shahkolaei *et al.*, November 2017)	**62.50**	**75.86**	**71.43**	**90.77**
MDQM	**78.57**	**82.35**	**82.93**	**96.77**

We can see that for almost all the metrics, the worn hole degradation has the highest value of F-measure. One reason is the fact that the worn holes distortion is one of the degradation prevailing in the degraded images of MHDID dataset. The other reason is that in the majority of images with worn holes degradation in the MHDID dataset, the amount of other degradations is negligible. Therefore, SVM classifier can detect this degradation category with more precision.

Table 5.5 lists the accuracy percentages for popular NR-IQA metrics on MHDID dataset. The results clearly indicate that the proposed metric has the best accuracy value for detecting degradation types in comparison with others.

Table 5.5 Accuracy comparison between six NR-IQA metrics and the proposed metric. Note that top two metrics are highlighted. Values are in percent.

NR Indices	DIIVINE	BLIINDS-II	BRISQUE	CurveletQA	BQMS	VDQAM	MDQM
Accuracy	47.76	49.25	55.22	49.25	34.32	**76.11**	**85.07**

Fig. 5.6 shows the histogram of run-time (seconds/image) and accuracy for seven blind IQA metrics. The blue and red bars demonstrate the run-time and accuracy, respectively. An efficient method should have a higher value of accuracy and lower value of run-time. It is clear from Fig. 5.6 that the proposed method has the best accuracy and its run-time is at an acceptable level in comparison with other NR-IQA metrics. It can be observed from Fig. 5.6 that the lowest run-time and highest accuracy belong to the BRISQUE and MDQM metrics, respectively, while the run-time efficiency of the BLIINDS-II metric is the worst.

5.4.3 Computational load analysis

Having illustrated that MDQM metric performs remarkably well in predicting the document image quality, now we demonstrate that this metric also has an acceptable complexity among other state-of-the-art metrics. Table 5.6 shows the number of features and run time of seven NR-IQA metrics when were applied to the images of size 1024×1280. The experiments were performed on a modern desktop computer (Core i7 3.40GHz CPU, 16 GB of RAM,

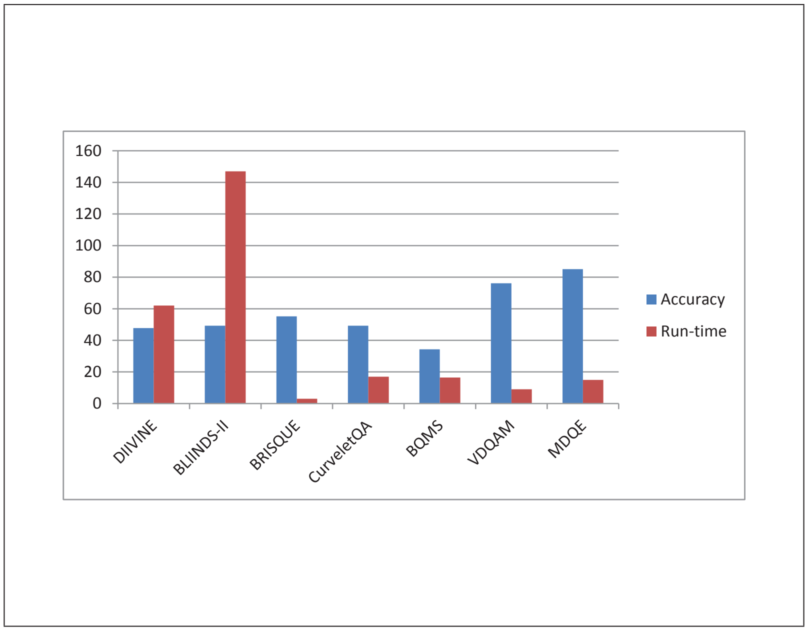


Figure 5.6 The histogram of run-time (seconds/image) and accuracy for seven NR-IQA metrics.

MATLAB 2015*b* and windows 7 Pro 64-bit). From the Table 5.6, it is obviously clear that the proposed metric ranks the third fastest metric in comparison with others. Although the number of features in MDQM is larger than others metrics, its complexity is still admissible and it is better than 4th to 7th metrics.

Table 5.6 Comparison of run-time (milliseconds/image) and the number of features for seven NR-IQA metrics.

NR-IQA metrics	no. of features	run time (ms)
1) BRISQUE (Mittal *et al.*, 2012)	36	3195
2) VDQAM (Shahkolaei *et al.*, November 2017)	40	8621
3) **MDQM**	124	11871
4) CurveletQA (Liu *et al.*, 2014)	12	13396
5) BQMS (Gu *et al.*, 2016)	13	13480
6) DIIVINE (Moorthy & Bovik, 2011)	88	61793
7) BLIINDS-II (Saad *et al.*, 2012)	24	148574

5.5 Conclusion

In this paper, a new blind image quality assessment metric (MDQM) is proposed to assess the quality of ancient document images. This metric works based on the extraction of some features from MSCN coefficients, MSCN of gradient information and local phase on the two layers and the entire degraded document image. According to the performance analysis of the seven no-reference image quality assessment metrics, the proposed metric achieved the highest correlation with the human judgments on the VDIQA and MHDID datasets. Also, a degradation modeling based on the proposed metric is defined to estimate the probability of different types of degradation, namely, paper translucency, stain, readers annotations and worn holes, in damaged manuscripts. The experimental results indicate that the proposed metric have the significant ability for degradation classification while it has a moderate complexity.

In the future work, we will further consider other existing physical noises in damaged manuscripts by proposing a new dataset which has a huge number of degraded images. Moreover, proposing the degradation modeling will be considered as another target in our future work.

Acknowledgment

The authors thank the NSERC of Canada for their financial support under Grants RGPIN 138344-14.

CHAPTER 6

ARTICLE IV- DEGRADED MEDIEVAL DOCUMENTS: DATASET AND QUALITY ASSESSMENT BASED ON A SALIENCY APPROACH

Atena Shahkolaei[1], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Submitted to Pattern Recognition Letters
June 2019

Abstract

This paper considers quality assessment of medieval document images with textual and pictorial content in the presence of physical degradations. Physical distortions have appeared on these documents through decades of deterioration and improper handling. These documents are of historical and cultural importance. For studying, preserving and archiving these documents, automatic methods are necessary given that a large amount of these documents is available. For better understanding of the physical distortions on document images, two contributions are provided in this work. First, a new dataset, called Degraded Medieval Document Dataset (DMDD), is introduced along with the subjective ratings on their quality for the first time in the literature. The DMDD dataset is freely accessible to researchers worldwide for research purposes. Second, we propose a no-reference metric, Blind Quality Assessment for DMDs (BQAD), in order to evaluate the quality of DMD images. The proposed metric is based on the fact that the human visual system (HVS) has different sensitivity to the pictorial and textual parts of a DMD. A new color saliency approach is used for segmentation of the DMD images into three layers. The extracted features from these layers are mapped to the subjective quality scores by a regression analysis. The proposed BQAD metric is the first attempt to assess the quality of DMD images. The experimental results performed on the DMDD dataset show that the proposed metric has a good correlation with the HVS with superior performance in comparison with the existing metrics in the literature. The developed dataset will be available at http://www.synchromedia.ca/system/files/DMDD.zip.

Keywords

Degraded medieval document images, Mean opinion score, Color saliency, Color gradient, Support vector regression

6.1 Introduction

Recent years have seen considerable growth in evaluation and enhancement of the ancient, physically distorted document images. DMD images are a subset of the historical document images and may contain both graphical and textual content. These documents belong to the period between 5^{th} and 15^{th} centuries. While a very large amount of these documents have survived, they might be severely degraded through time and improper handing.

Several researches have addressed the automatic processing of the DMD images for tasks like segmentation (Baechler & Ingold, 2011; Alberti *et al.*, 2019), transcription (Fischer *et al.*, 2009), etc. In the literature, enhancement of these documents is considered based on advanced imaging systems (Knox *et al.*, 2008; Goltz *et al.*, 2010; Montani *et al.*, 2012; Pottier *et al.*, 2019). Usually, these techniques are used to help in reading highly degraded medieval images. Recent studies have shown that although single-image enhancement of the highly degraded documents is very challenging, it is still possible to automatically assess their quality (Shahkolaei *et al.*, November 2017; Atena Shahkolaei, 2019). In (Shahkolaei *et al.*, November 2017), potential applications of the document image quality assessment for physically degraded documents are listed. Additional information on the degradation types is also of high interest (Atena Shahkolaei, 2019).

Dealing with the physically degraded documents, it is obvious that original pristine-quality document images are not available. Therefore, their image quality assessment (IQA) is in the category of no-reference IQA (NR-IQA). Usually, NR-IQA models are trained on the available datasets. There are recent efforts in developing datasets of degraded documents along with the human judgments (Shahkolaei *et al.*, November 2017; A. Shahkolaei, A. Beghdadi, S. Al-maadeed and M. Cheriet, 2018; Atena Shahkolaei, 2019). In the literature, different blind quality assessment metrics were proposed for evaluation of screen content images (SCI), Natural Scene Images (NSI) and Historical Document Images (HDI). In the following, we provide an overview on the metrics in each category.

In (Gu *et al.*, 2016), a no-reference image quality assessment (NR-IQA) metric which is called blind quality measure for SCIs (BQMS) was proposed. This metric extracts 13 features from the SCIs based on the free energy principle. Yummy et al. (Fang *et al.*, 2018) proposed an NR metric based on the statistical luminance and texture features (NRLT) for SCIs. Anish Mittal et al. (Mittal *et al.*, 2013) proposed an NR-IQA metric for quality evaluation of NSI, namely Natural Image Quality Evaluator (NIQE). This metric is based on the creation of a 'quality aware' collection of statistical characteristics based on a simple and successful space domain natural scene statistics (NSS) model. The Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) proposed by (Moorthy & Bovik, 2011). This metric first classifies distortion types. Then, subband coefficients of discrete wavelet transform (DWT) are fitted by generalized Gaussian distribution (GGD). Recently, two blind image quality assessment methods were proposed for evaluation of the HDIs. The Visual Document image Quality Assessment Metric (VDQAM) works based on the extracted statistical features from four layers of the document image (Shahkolaei *et al.*, November 2017). Shahkolaei et al. (Atena Shahkolaei, 2019) proposed a blind quality assessment metric, namely Multi-distortion Document Quality Measure (MDQM). This metric works on the two layers of degraded documents. Features are extracted from the gradient, MSCN coefficients and mean of local phase.

It is possible to propose better performing quality assessment metrics for specific documents like medieval documents. In this paper, salient pixels of medieval documents are segmented and used to propose a quality metric. Our assumption is that medieval documents contain pictorial parts in color. First, a dataset of medieval documents is created with the human ratings on their quality. Then, we propose a new no-reference image quality assessment (BQAD) metric that works based on the extraction of statistical features from three segmented layers of the DMD images, namely, text, non-text and graphics. For segmentation, color saliency methods are utilized. The experimental results show that the proposed metric provide quality scores that are fairly correlated with those given by the human visual system.

In this paper, we first provide a detailed description of the developed dataset. In section 6.3, we discuss the proposed metric. In section 6.4, experimental results are provided. Finally, section 7.3 provides a conclusion.

6.2 DMDD dataset

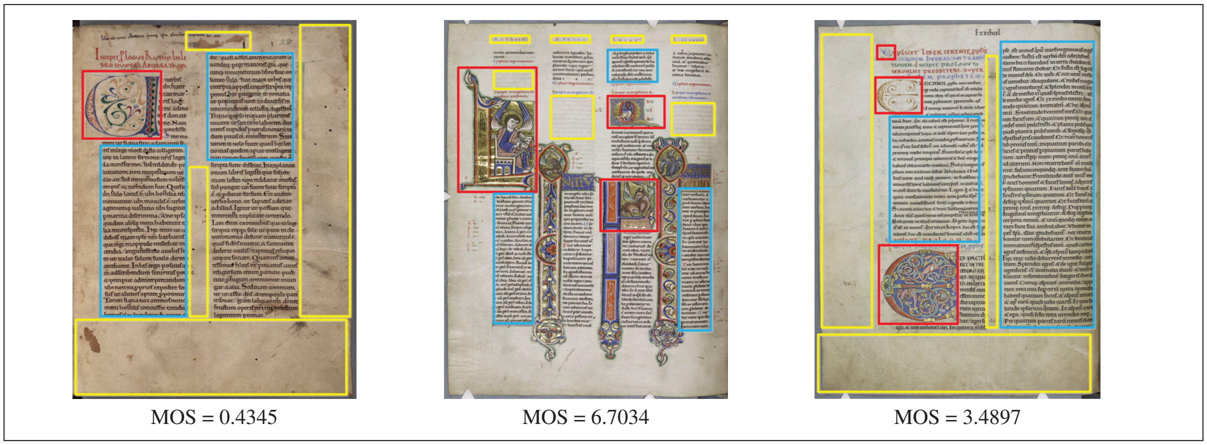
In this section, every step involved in the creation of the DMDD dataset is explained.

6.2.1 Selection of degraded images

The content of the selected images in a new dataset may inflict varying effects on IQA metrics (Winkler, 2012). Therefore, the selection of images is an important part of the creation of a new dataset. Totally, 150 degraded document images with diverse original content that contain both textual and graphical regions are selected from the Parker library of Stanford University. These document images are selected from 11^{th}, 12^{th}, and 13^{th} centuries. The authors used different colors, shapes and decorations for organization of manuscripts which help readers for better understanding. There are three categories of graphics in these documents: i) Litterae Notabiliores, ii) Capital, and iii) Rubric. Three examples with partial segmentation are shown in Fig. 6.1.

6.2.2 Subjective evaluation

Fifteen students between 20-35 years old were employed to judge between two DMD images in each iteration. A briefing session was held for subjects in order to explain the comparison criteria. The subjects were asked to judge between two DMD images based on the overall quality of degraded documents.



MOS = 0.4345

MOS = 6.7034

MOS = 3.4897

Figure 6.1 Three DMD images with partial segmentation. Text, non-text and graphical regions are shown with blue, yellow and red colors, respectively.

The pair comparison rating (PCR) method (Sun *et al.*, 2017) is a simple yet effective subjective evaluation method. Each subject must choose one of the three options which were shown in each trial: “The quality of the left image is better", “The quality of the right image is better" and “There is no difference between two DMD images". The scores of 1, -1 are assigned to the image with higher and lower quality, respectively. If the subject could not choose the document with better quality, a score of 0 is considered in this case.

The experiments were done in the normal condition in the different laboratories of ETS University of Canada. All desktops had a 64-bit Windows operating system and 32GB of RAM. More detailed information on the DMDD dataset is provided in Table 6.1.

6.2.3 Data processing

Based on the subjective pair comparisons, the judgments of three subjects were removed as being outliers which account for 20% of the subjects. The rest of the ratings given by the remaining 12 subjects were normalized between 0 and 9:

$$z_i = 9 \times \frac{x_i - min(x)}{max(x) - min(x)} \tag{6.1}$$

Table 6.1 Detailed information about the DMDD dataset. MTD stands for Multiple Types of Distortions.

Dataset	DMDD
Number of images	150
Contain color/graphics	Yes
Number of subjects	15
Number of outliers	3
Number of ratings per each image pair	10
Image format	JPG
Size of screen	17 inch
Size of images	1280 × 1024
Screen resolution	1600 × 1200
Range of scores	[0 9]
Type of dataset	MTD
Total size of dataset	˜36 MB
Data format	MOS

where $x = (x_1, ..., x_n)$ and z_i is the i^{th} normalized data. It should be mentioned that the quality score of 0 and 9 correspond to the lowest and highest perceptual quality, respectively. Please refer to the Fig. 6.1 shows three DMD images with associated MOS values.

6.3 Proposed metric (BQAD)

In this section, the proposed opinion-aware blind image quality assessment metric is described. The BQAD metric works on the three segmented layers of each DMD image: text, non-text and graphics. The reason for this segmentation is that these layers do not have an equal impact on the HVS judgments. Fig. 6.1 shows three selected DMD images from the DMDD dataset with partial segmentation that are highlighted with three distinct colors.

6.3.1 Segmentation

As mentioned earlier, the sensitively of HVS to various parts of DMD images is not equal, due to the existing different degradations, text, non-text and graphical areas in these documents. Therefore, it is worth to segment each DMD image into different layers and consider the statistical features of each layer, separately. In this section, each input DMD image is segmented into three layers based on the BQAD method. More details about the layer's segmentation and proposed saliency method are provided concerning the following subsections.

6.3.1.1 Color/graphical layer segmentation

In the DMDD dataset, graphics are written in color. It is common to use classical approaches such as the Gaussian mixture model with expectation maximization and K-Means for color image segmentation. However, these methods have shown poor segmentation results on the images of the DMDD dataset because images in the DMDD dataset are degraded, the background color might be very similar to the colored text, and the number of classes varies from one image to another. Therefore, in this work, possible graphical pixels are extracted based on a gradient color saliency and a standard deviation (STD) based color saliency.

In (Itti & Koch, 2001), it is shown that color is important to express saliency. In (van de Weijer *et al.*, 2006), color distinctiveness is explicitly incorporated into the design of salient point/edge detection. In other words, the goal is to focus on the more distinctive points. The transformation of the image to achieve this goal is called color saleincy boosting. Color boosting transformation is done in the opponent color space:

$$\begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{6.2}$$

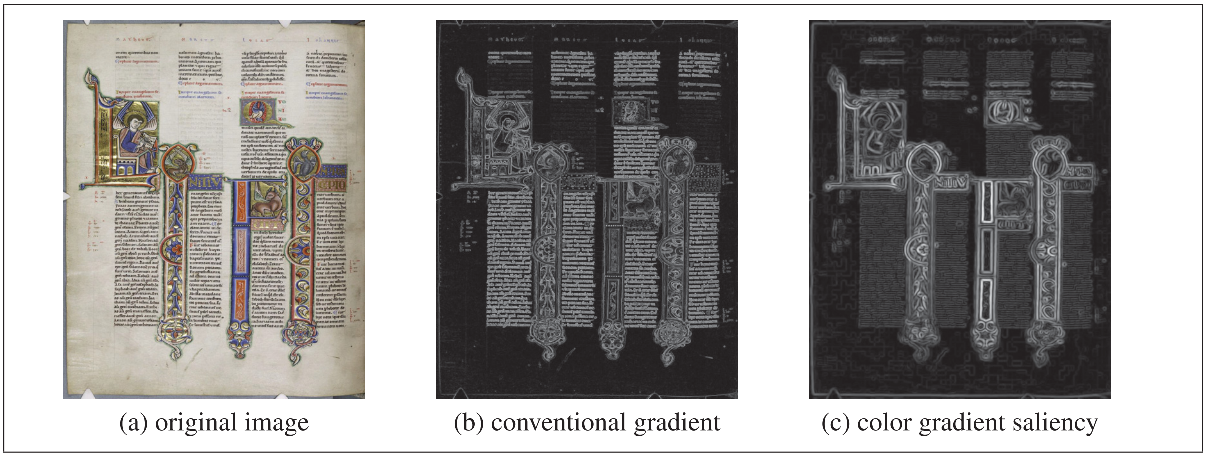
where R, G and B refer to the color image channels. The color boosting is a weighting of the individual opponent channels:

$$\begin{pmatrix} 0.850 \times o_1 \\ 0.524 \times o_2 \\ 0.065 \times o_3 \end{pmatrix} \tag{6.3}$$

where the sum of the squared weights is equal to 1. Let o_x and o_y denote the horizontal and vertical gradients of an opponent channel. Then, the color gradient saliency g is computed by:

$$g = \sqrt{o_{1x}^2 + o_{1y}^2 + o_{2x}^2 + o_{2y}^2 + o_{3x}^2 + o_{3y}^2} \tag{6.4}$$

Fig. 6.2 shows the difference between the color gradient saliency and the conventional gradient magnitude with Sobel operator (I. Sobel, 1968). In Fig. 6.2(c), only color edges are detected which is useful for segmenting graphics in the DMD images.


(a) original image

(b) conventional gradient

(c) color gradient saliency

Figure 6.2 An original DMD image (*a*), its grayscale edge magnitude with Sobel operator (*b*) and color saliency based edge map (*c*).

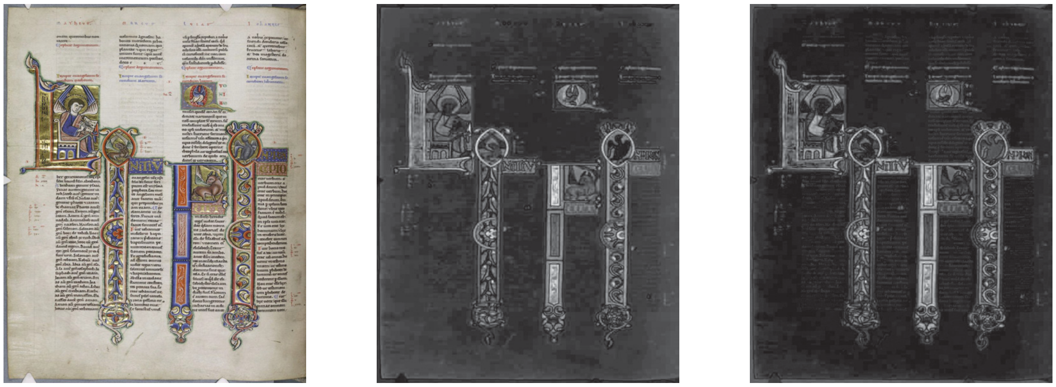
Since background often has less gradient information, it can be distinguished from the texts with the same color as background. This approach, however, cannot always distinguish between non-salient text and the salient text (colored). Therefore, we use a simple and efficient color saliency method to better classify image pixels into non-salient and salient color pixels.

Given input color image I, its mean image μ and standard deviation image σ are computed by:

$$\mu(i,j) = \frac{1}{3} \sum_{c=R,G,B} \mathrm{I}_c(i,j) \tag{6.5}$$

$$\sigma(i,j) = \left[\frac{1}{2} \sum_{c=R,G,B} \left|\mathrm{I}_c(i,j) - \mu(i,j)\right|^2\right]^{1/2} \tag{6.6}$$

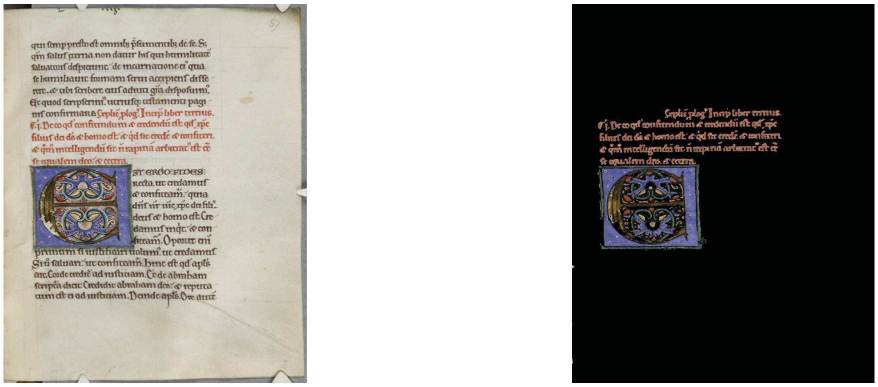
The proposed method uses two σ images, the first one is calculated from the input color image I (σ_1) and the other one from the variance normalized I (σ_2). Fig. 6.3 shows example output of σ_1 and σ_2.



(a) original image (b) σ_1 (c) σ_2

Figure 6.3 An original DMD image (a), its standard deviation image (b) and standard deviation of normalized image (c).

Given the color gradient saliency, σ_1 and σ_2, thresholding and morphological operations are used to extract the graphical parts of the DMD images. Fig. 6.4 shows an example of the color segmentation results.

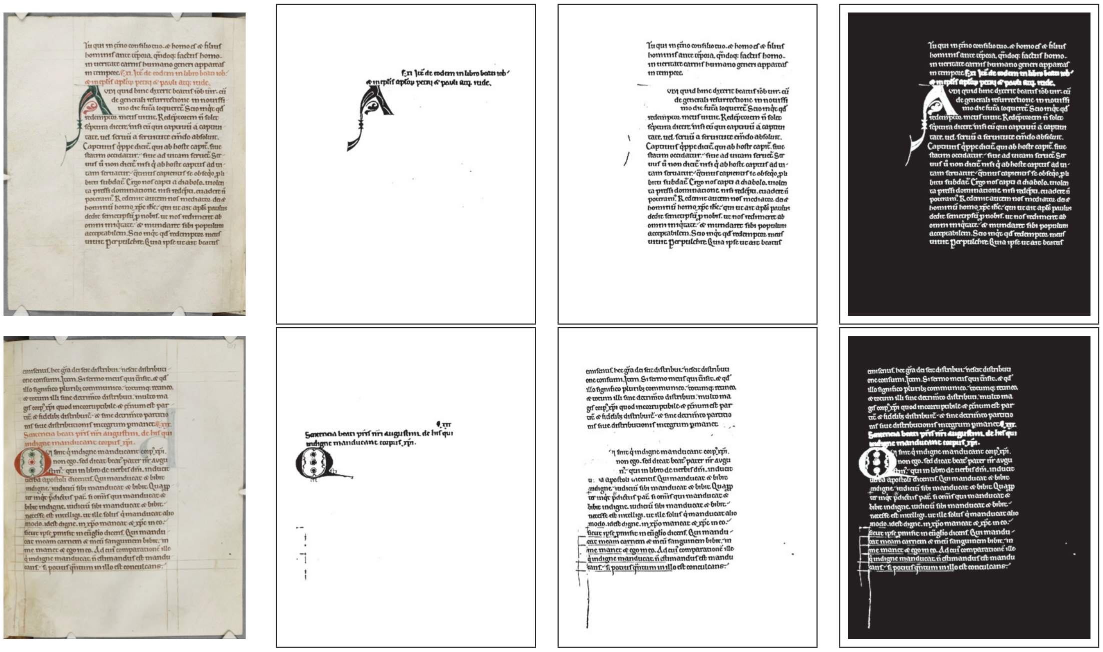


(a) original image (b) graphic/color layer

Figure 6.4 Color segmentation results

6.3.1.2 Background and text segmentation

We also use the phase-based binarization method of (Nafchi *et al.*, 2014) to segment the non-graphic text from the background. These two layers as well as the graphics layer constitute the three layers that will be used in the proposed DIQA model. Fig. 6.5 shows two examples of the proposed segmentation into three layers. Note that the segmentation results might not be perfect but satisfactory enough for the quality assessment task.



(a) original image (b) color layer (c) text layer (d) non-text layer

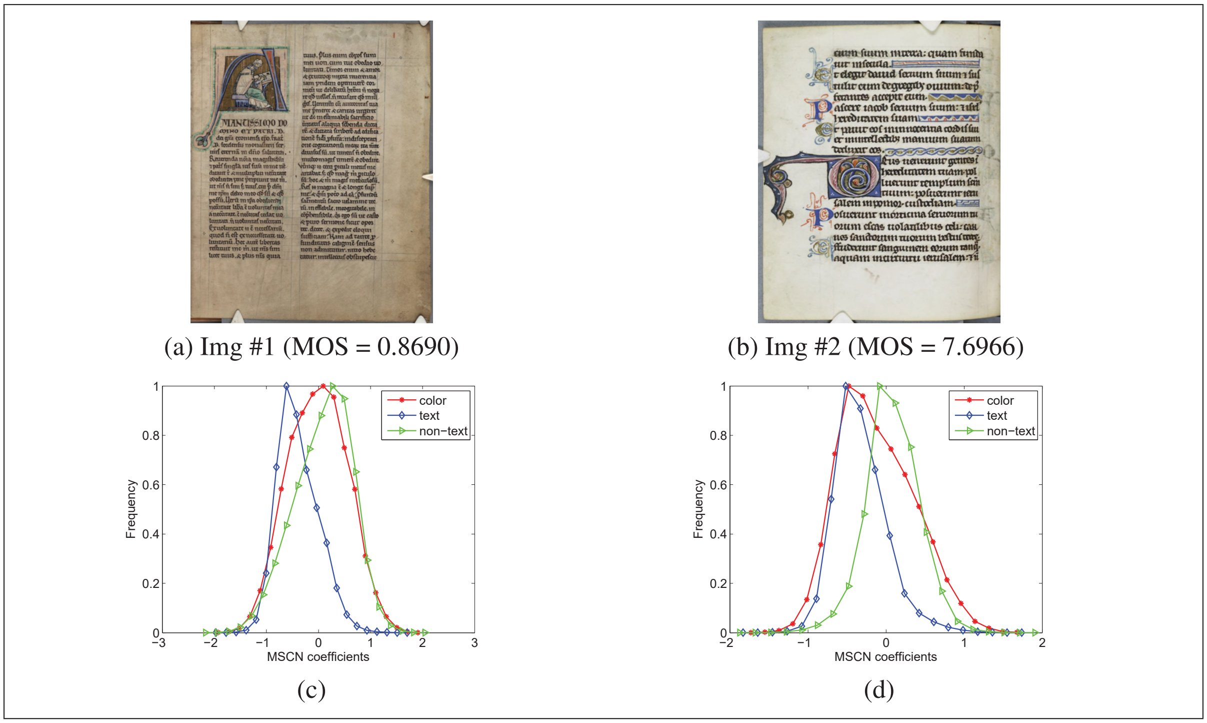
Figure 6.5 Results of segmentation into three layers of graphics, text and non-text.

6.3.2 MSCN coefficients

Conversion to the MSCN coefficients is a common decorrelation technique to remove dependency of the features in images. The MSCN coefficients tends a Gaussian distribution for a noise-free image. In this work, statistics of the MSCN coefficients are computed from the segmented layers of the DMD images. Assume that I is an original DMD image, MSCN coefficients is computed as the following:

$$\mathrm{MSCN}(i,j) = \frac{\mathrm{I}(i,j) - \mu(i,j)}{\sigma(i,j) + 1} \tag{6.7}$$

where $\mu(i,j)$ and $\sigma(i,j)$ are the local mean and standard deviation respectively (Atena Shahkolaei, 2019). Fig. 6.6 shows the joint histograms of MSCN coefficients for three extracted layers of two selected DMD images from the DMDD dataset. It is clear that these histograms tend to a Gaussian distribution. Due to the different behavior of the MSCN distributions for each layer, the statistics of each layer can contribute to the overall quality score that is predicted by the BQAD metric.



(a) Img #1 (MOS = 0.8690)

(b) Img #2 (MOS = 7.6966)

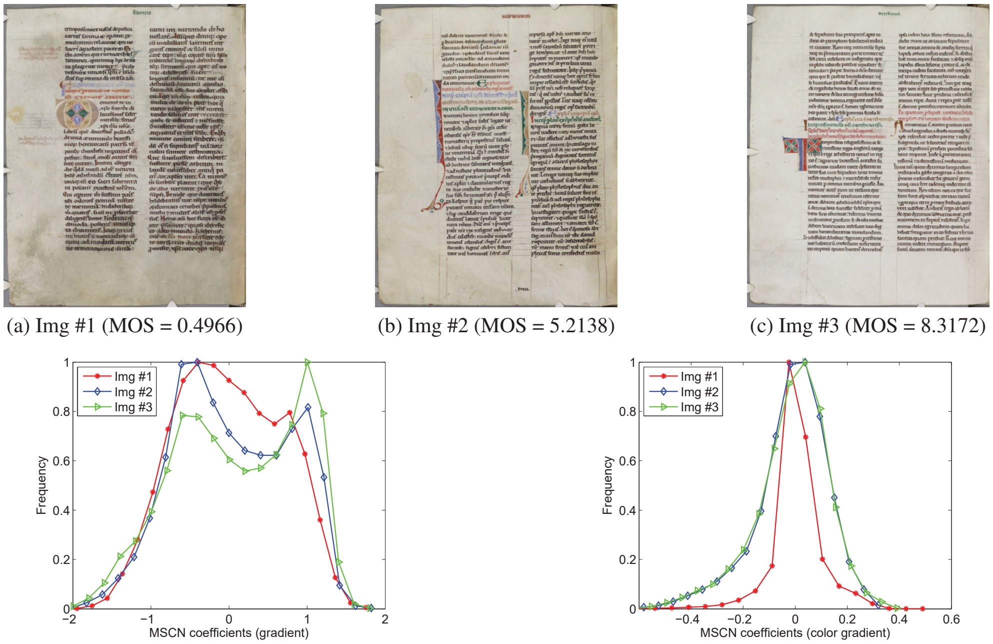
(c)

(d)

Figure 6.6 Histograms of MSCN coefficients for three layers of the two DMD images with different MOS values. (c) MSCN distribution of three layers in Img #1; (d) MSCN distribution of three layers in Img #2.

6.3.3 Color gradient statistics

Color gradient saliency is not only used for DMD image segmentation, its MSCN statistics on the color layer are also used for quality assessment. Fig. 6.7 shows three examples of the DMD images along with the MSCN coefficients of the conventional gradient and color gradient saliency. It can observed from the plots that MSCN coefficients of the color gradient better follow the Gaussian distribution than that of the conventional gradient.



(a) Img #1 (MOS = 0.4966)

(b) Img #2 (MOS = 5.2138)

(c) Img #3 (MOS = 8.3172)

Figure 6.7 Comparing MSCN of gradient histograms vs MSCN of color gradient saliency histograms for three DMD images.

6.3.4 Feature extraction

MSCN coefficients and MSCN of color gradient are fitted with the generalized Gaussian distribution (GGD). The zero mean GGD is defined as:

$$f(x;\varphi,\sigma^2) = \frac{\varphi}{2\kappa\Gamma(1/\varphi)} e^{\left(-\frac{|x|}{\kappa}\right)^{\varphi}} \tag{6.8}$$

where φ and κ are effective quality aware features that are estimated based on the moment-matching approach. The variance of the distribution is shown with σ^2. Gamma function $\Gamma(.)$ is defined as:

$$\Gamma(\varphi) = \int_0^\infty t^{\varphi-1} e^{-t} dt \ , \qquad \varphi > 0 \tag{6.9}$$

When the shape of the histogram is not symmetric, asymmetric generalized Gaussian distribution (AGGD) is more accurate (Mittal *et al.*, 2012). The AGGD with zero-mode is defined by:

$$f(x;\vartheta,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{\vartheta}{(\sigma_l+\sigma_r)\Gamma(1/\vartheta)} \exp\Big(-(\frac{-x}{\sigma_l})^{\vartheta}\Big), & x < 0 \\ \frac{\vartheta}{(\sigma_l+\sigma_r)\Gamma(1/\vartheta)} \exp\Big(-(\frac{x}{\sigma_r})^{\vartheta}\Big), & x \geq 0 \end{cases} \tag{6.10}$$

The parameters $(\vartheta,\sigma_l^2,\sigma_r^2)$ are also important quality aware features.

Given the MSCN coefficients of the DMD image, the two parameters of GGD (φ, σ^2) and three parameters of AGGD $(\vartheta,\sigma_l^2,\sigma_r^2)$ for each of the three segmented layers constitute the first part of the feature vector in this work. Given the MSCN coefficients of the color gradient saliency, the five parameters of GGD and AGGD on the color layer of the DMD image constitute the second part of the feature vector of the proposed metric. All extracted features are computed in two scales. The number of 40 features (20 features for each scale) are extracted from each DMD image. Finally, these features are mapped to the quality scores by support vector regression (SVR).

6.4 Experimental results

In this section, we compare the performance of the BQAD metric with some popular metrics which were proposed for NSI, SCI and HDI images. In order to evaluate the performance of these metrics, the developed DMDD dataset is used.

The performance of the BQAD metric was evaluated with the Spearman rank correlation coefficient (SRC) for measurement of the monotonicity; the Pearson linear correlation coefficient (PCC) for measurement of the linearity and the Kendall rank correlation coefficient (KRC) for measurement of the similarity between subjective rating and quality scores. In order to remove the non-linearity between quality scores and MOS values, the following regression function is used:

$$f(x) = \alpha_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\alpha_2 (x - \alpha_3)}} \right) + \alpha_4 x + \alpha_5 \tag{6.11}$$

where α_i, $i = 1, 2, ..., 5$ are regression model parameters.

Table 6.2 Performance comparison of the BQAD metric with six IQA algorithms on the DMDD dataset for three cases: 20% train and 80% test, 50% train and 50% test, 80% train and 20% test.

		SRC	PCC	KRC
NR-metric	Type	20%-80%		
NIQE	NSI	0.3959	0.4191	0.2732
DIIVINE	NSI	0.5714	0.6011	0.3989
BQMS	SCI	0.3791	0.4720	0.2618
NRLT	SCI	0.6084	0.6260	0.4183
VDQAM	HDI	0.6092	0.6357	0.4201
MDQM	HDI	0.6311	0.6467	0.4414
BQAD	DMD	**0.6700**	**0.7043**	**0.4815**
50%-50%				
NIQE	NSI	0.4094	0.4474	0.2848
DIIVINE	NSI	0.6788	0.7054	0.4981
BQMS	SCI	0.4269	0.5304	0.3025
NRLT	SCI	0.6667	0.7318	0.4994
VDQAM	HDI	0.6875	0.7363	0.5008
MDQM	HDI	0.7013	0.7435	0.5301
BQAD	DMD	**0.7573**	**0.7863**	**0.5639**
80%-20%				
NIQE	NSI	0.4117	0.4752	0.2937
DIIVINE	NSI	0.7110	0.7613	0.5241
BQMS	SCI	0.4623	0.5829	0.3330
NRLT	SCI	0.6877	0.7700	0.4954
VDQAM	HDI	0.7181	0.7788	0.5337
MDQM	HDI	0.7387	0.8010	0.5623
BQAD	DMD	**0.8095**	**0.8474**	**0.6136**

In Table. 6.2, the performance of the proposed metric and other metrics on the DMDD dataset are evaluated. In order to evaluate the generalization ability of the evaluated metrics, the dataset is divided into the three randomly chosen subsets: 20% train 80% test, 50% train 50% test, and 80% train 20% test. For these three cases, the median value of 1000 times train-test is reported in Table. 6.2. As expected, the performance of our metric is remarkably better than the rest of the metrics for different randomly chosen subsets of training and testing. From Table. 6.2 it can be observed that the performance of metrics VDQAM and MDQM that were proposed to assess HDI images is closer to the results of the BQAD metric than the other metrics.

The correlation between each segmented layer of DMD image in DMDD dataset and MOS values are reported in Table 6.3. These results are also reported based on the median values of 1000 times train and test for three cases: 20% train and 80% test, 50% train and 50% test, 80% train and 20% test. It can be observed that layer L_1 attain a higher correlation with MOS values in comparison with the two other layers. From this table, we can conclude that the HVS is more sensitive to the L_1, L_2 and L_3 layers according to the proposed metric, respectively.

Table 6.3 Correlation between three segmented layers of color (L_1), text (L_2) and non-text (L_3) with the human judgments.

		L_1	L_2	L_3
20%-80%	SRC	**0.6286**	0.5579	0.2938
	PCC	**0.6892**	0.6224	0.4069
	KRC	**0.4488**	0.3836	0.1972
50%-50%	SRC	**0.6327**	0.6045	0.3017
	PCC	**0.7194**	0.7005	0.4838
	KRC	**0.4599**	0.4255	0.2672
80%-20%	SRC	**0.6176**	0.5929	0.4075
	PCC	**0.7413**	0.7263	0.5285
	KRC	**0.4527**	0.4231	0.2794

6.4.1 Computational complexity

The feature extraction time for seven blind image quality assessment metrics was calculated in Fig. 6.8. This experiment was done on a PC with 32GB of RAM and 64-bit Windows operating system. The software platform was Matlab R2015b. The size of the tested image was 1024 × 1280. It can be seen from this figure that the proposed metric has a moderate computational complexity.

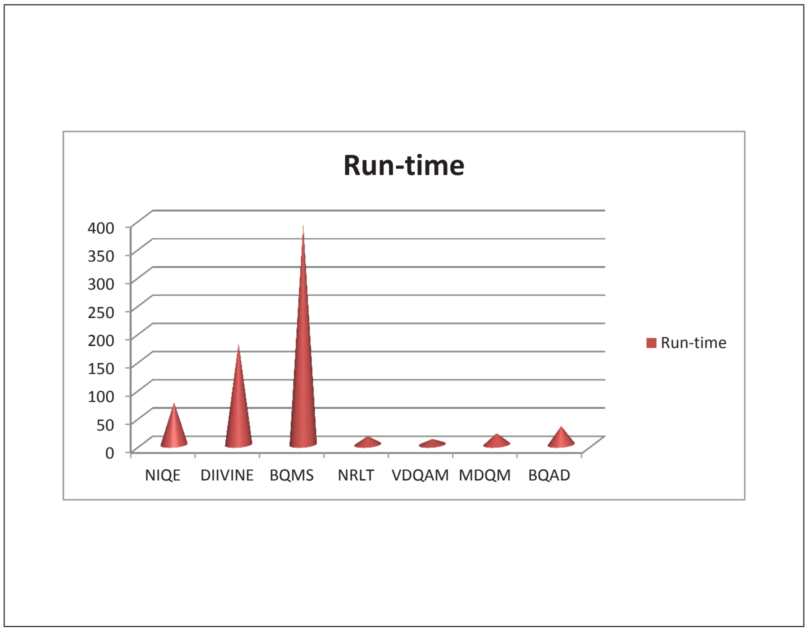


Figure 6.8 The chart of feature extraction time (in seconds) for seven IQA methods.

6.5 Conclusion

In this paper, we proposed a new dataset (DMDD) and a new NR-IQA metric for the evaluation of the DMD images. The DMDD dataset contains 150 degraded documents which have both of the textual and pictorial regions. The BQAD metric works on the three layers of each DMD image that are segmented by using a color saliency approach. The experimental results demonstrate that the performance of the BQAD metric is surprisingly better than the state-of-the-art metrics. Although the paper presents a new blind quality metric and a new dataset for DMD images for the first time and the performance of the proposed metric on the DMDD dataset is remarkably good, there is scope for improving certain aspects of the algorithm. In particular, the number of DMD images for the creation of a new dataset can be increased. Also, investigating and proposing a new blind metric which can improve the results of our method, is another direction for future research.

Acknowledgment

The authors thank the NSERC of Canada for their financial support under Grants RGPIN 138344-14.

CHAPTER 7

CRITICAL DISCUSSION OF SOME CONCEPTS OF THE THESIS

This thesis has addressed a number of problems related to the document image quality assessment and degradation classification. The introduction and literature review (Chapter 1) showed challenges associated with processing physically distorted document images and limitations of the existing metrics. Based on these challenges and limitations, we established four research objectives in Chapter 2 that led to the proposition of three NR-DIQA metrics, a degradation classification method, and three datasets with associated subjective ratings. These datasets, metrics and methods were presented, discussed and evaluated in Chapter 3, Chapter 4, Chapter 5, and Chapter 6. Our contributions are now discussed in the following sections by considering their advances made in the state of the art document image quality assessment, with a focus on their strength and limitations.

7.1 Subjective and objective quality assessment of degraded document images (RQ1)

To the best of our knowledge, there is no dataset with associated human ratings or any quality assessment metric available for physically degraded document images. This has led to our attempt to develop a dataset and a metric with a high correlation with the human visual system. In the design of the proposed NR-DIQA metric, a multi-layer approach is used. It includes the extraction of the statistical features in the spatial domain from each layer of the document images. The motivation to use a multi-layer approach was to follow HVS that may give different weights to the text and non-text in a document image. One drawback of the proposed metric is its high complexity because features are extracted from four layers. Also, empty layers may lead to undesirable results. Another problem with the proposed metric is that segmentation into four layers is not accurate.

7.2 Blind quality assessment metric and degradation classification for degraded document images (RQ2)

This work offers three modifications over the previous work. First, a larger dataset of physically distorted documents is developed that has the advantage of having four labeled types of distortion. A potential flaw of the developed dataset is the number of degraded images and types of degradations. Indeed, it is better that the number of degraded documents and types of degradations be increased in the development of a new dataset in the future. Second, a more efficient NR-DIQA metric is proposed by using a new set of features in just two layers. Finally, the extracted features are used to classify degradations into one of the four labeled distortion types. This means that our work provides complementary information on both severity and type of distortions. As mentioned before, the proposed model cannot detect all existing noises in ancient documents. Therefore, this model can be further improved to identify more types of physical distortions in HDIs. Introducing an automatic degradation modeling can help in better understanding and analysis of document image quality and physical degradation.

7.3 Subjective and objective quality assessment of DMD images based on a saliency approach (RQ3)

Customizing the NR-DIQA metrics to assess specific type of the document images shows advantages over general purpose NR-DIQA metrics. For example, in our last work, a dataset (called DMDD) and a blind quality metric were proposed for degraded medieval documents (DMDs). DMDD dataset was introduced along with the subjective ratings for the first time in the literature. The proposed metric was based on the fact that HVS has different sensitivity to the pictorial and textual parts of a DMD. A color saliency approach and a phase-based binarization method were used for segmentation of the DMD images into three layers. The extracted features from these layers were mapped to the subjective quality scores by regression analysis. We have trained and tested the proposed metric on DMDD dataset and shown their usefulness. The weakness of the proposed metric is that it necessarily needs color document images.

CONCLUSION AND RECOMMENDATIONS

8.1 General conclusion and future work

In this work we have presented original contributions to the state of the art in the field of document image quality assessment. Overall in this thesis, three datasets, three quality assessment metrics and a degradation classification model for historical document images were proposed. The proposed metrics and methods achieved a superior performance over the state of the art methods. Our work has led to a number of publications in international journals and a conference that have shown the usefulness of our work.

Physical distortions in document images can still be assessed even if they cannot be removed. In this thesis, we have addressed the problem of quality assessment of physically distorted document images for the first time. For quality assessment of ancient document images, the creation of datasets associated with MOS values is necessary. Therefore, three datasets were built based on the human ratings for HDI and DMD images in our work. We then proposed three document image quality assessment metrics through a unified approach. Our approach was to segment document images into different layers following the document characteristics and human visual system behavior. We have also presented a degradation classification model to estimate the probabilities of four common types of physical distortions.

No-reference image quality assessment metrics are of exceptionally high interest because in real-world applications, usually, the original signal is not available. The quality assessment of old documents is an important issue that has not been tackled in state of the art. We proposed a dataset and a blind quality evaluation metric for historical documents in the first research work. In the proposed metric, each degraded document image is segmented into four layers according to the log-Gabor filters, and the spatial statistics of these layers are used for quality assessment. We have tested the state-of-the-art no-reference image quality assessment metrics on the developed dataset, and we were able to achieve the highest performance. We believe that this contribution of the thesis can motivate the researchers to improve the results by proposing better performing metrics.

In our other work, a dataset with four categories of distortion types, a blind quality assessment metric and a degradation classification model were presented. The developed dataset contains 335 historical document images which are classified into four categories based on their distortion types, namely, paper translucency, stain, readers' annotations and worn holes. The proposed metric is based on three sets of spatial and frequency image features. These features were extracted from two layers of text and non-text and mapped to the MOS values by a regression function. As mentioned before, a degradation classification model was provided to detect and estimate different types of degradations in degraded document images. Proposing an automatic degradation modeling which can identify and determine more types of distortions in ancient documents is recommended for future works. Indeed, this contribution of the thesis can open a path for proposing automatic multi-purpose degradation models.

The last, a dataset and a blind quality metric were proposed for degraded medieval documents. DMDD dataset was introduced along with the subjective ratings for the first time in the literature. The proposed metric was based on the fact that HVS has different sensitivity to the pictorial and textual parts of a DMD. A color saliency approach and a phase-based binarization method were used for segmentation of the DMD images into three layers. The extracted features from these layers were mapped to the subjective quality scores by regression analysis. We have trained and tested the proposed metric on DMDD dataset and shown their usefulness.

Future work

Historical document images suffer from several types of distortions that accumulate and evolve over time. In our work, the proposed degradation classification model just detect and estimate the type and severity of four common degradations: paper translucency, stains , readers' annotations and worn holes. It would be interesting to propose an automatic degradation modeling that can detect several types of degradations in ancient documents by collecting sufficient training samples from each degradation.

As mentioned before, the proposed datasets in my research work are the first developed datasets for historical and degraded medieval documents. The number of degraded images in these datasets can be increased. In the future work, we will further consider other existing physical noises in damaged manuscripts by proposing a new dataset which has a huge number of degraded document images.

The proposed quality metrics in this work trained and tested on developed datasets (VDIQA, MHDID and DMDD), which include scripts Arabic. To the best of our knowledge, there is no dataset containing a multitude of writing scripts (e.g. Latin, Chinese, Cyrillic, Greek, etc.) in the literature which was distorted after years with associated MOS values. Therefore, the lack of existing a dataset with a multitude of writing scripts exist. It would be interesting to develop such datasets. Therefore, this issue can be considered as another important research topic for future work.

Extension of the proposed quality metrics for HDI and DMD images to increase the correlation with the HVS is of high interest. Although the correlation of the proposed metrics with HVS is remarkably high in our proposed metrics, there is still room for improvements.

It is not deniable that historical document images are commonly suffered from noises and degradations that their protection is really significant for preserving the civilization of each country. Therefore, enhancement methods can be designed to improve and preserve the quality of these documents. Indeed, enhancement methods can be proposed for each degradation in HDIs after detecting of these distortions in ancient documents.

All of these challenges and suggestions are mentioned in order to improve the present current works. We hope that it will inspire future research in these scientific areas.

8.2 Summary of contributions

The highlight of the major contributions of this thesis are:

- Analyzing statistics of degraded document images and introducing useful image features for document quality assessment in general,
- Developing two datasets of physically distorted documents with associated quality scores given by subjective evaluation,
- Proposing a degradation classification algorithm to estimate the probability of different common types of degradation in old document images, namely, paper translucency, stain, readers annotations and worn holes,
- Proposing two blind quality assessment metrics for degraded document images. These blind metrics evaluate the quality of degraded documents based on the extraction of some statistical features from different layers of historical documents,
- Proposing new phase-derived features for document image quality assessment that follow Gaussian distribution,
- Creating a new dataset based on the human judgments for degraded medieval documents for the first time in the literature,
- Development of a general quality assessment model for degraded medieval document images based on a saliency-based segmentation approach.

8.3 Articles in peer-reviewed journals and conferences

1. Atena Shahkolaei, Mohamed Cheriet: Degraded Medieval Documents: Dataset and Quality Assessment Based on a Saliency Approach, In Pattern Recognition Letters, Under review, July 2019.

2. Atena Shahkolaei, Azeddine Beghdadi, Mohamed Cheriet: Blind Quality Assessment Metric and Degradation Classification for Degraded Document Images, In Signal Processing: Image Communication, vol. 76, pp. 11-21, (2019).

3. Atena Shahkolaei, Hossein Ziaei Nafchi, Somaya Al-Maadeed, Mohamed Cheriet: Subjective and objective quality assessment of degraded document images, In Journal of Cultural Heritage, vol. 30, pp. 199-209, (2018).

4. Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam and Mohamed Cheriet: Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator. IEEE Access, vol. 4, pp. 5579-5590, (2016).

5. Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam and Mohamed Cheriet: MUG: A Parameterless No-Reference JPEG Quality Evaluator Robust to Block Size and Misalignment. IEEE Signal Processing Letters, vol. 23, no. 11, pp. 1577-1581, (2016).

6. Hossein Ziaei Nafchi, Atena Shahkolaei, Reza Farrahi Moghaddam and Mohamed Cheriet: FSITM: A Feature Similarity Index For Tone-Mapped Images. IEEE Signal Processing Letters, vol. 22, no. 8, pp. 1026-1029, (2015).

7. Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, Mohamed Cheriet: CorrC2G: Color to Gray Conversion by Correlation. IEEE Signal Process. Lett. 24(11): 1651-1655 (2017).

Articles in peer-reviewed conference proceeding

1. Atena Shahkolaei, Azeddine Beghdadi, Somaya Al-Maadeed, Mohamed Cheriet: MH-DID: A Multi-distortion Historical Document Image Database, In International Workshop on Arabic and derived Script Analysis and Recognition (ASAR 2018), Pages 156-160, (2018).

8.4 Other scientific activities

- ICIP 2015 (International Conference on Image Processing)- Helped in coordinating both the oral and poster sessions.
- CONGRÈS DES ÉTUDIANTS CHERCHEURS À L'ÉTS (CAEC-ETS) 2017- I had a poster about my research in this event. I also gave talks on the quality assessment of historical document images.
- ICIAR 2017 (International Conference on Image Analysis and Recognition)- Helped in coordinating the poster session and registration desk.
- CAN-CWIC 2017 (The ACM Canadian Celebration of Women in Computing)- I co-organized and helped in running the registration desk.

Paper reviewing

- IEEE Transactions on Image Processing (1 paper)
- International Conference on Image Processing (1 paper)
- International Conference on Document Analysis and Recognition (1 paper)

Award

École de technologie supérieure (ÉTS), Internal Scholarship (2018).

BIBLIOGRAPHY

A. Chetouani,A. Beghdadi, M. D. (2010, Pages: 714-717). Statistical Modeling of Image Degradation Based on Quality Metrics. *ICPR*.

A. Shahkolaei, A. Beghdadi, S. Al-maadeed and M. Cheriet. (2018). A Multi-Distortion Document Image Dataset Towards No-Reference Quality Assessment. *2nd IEEE Int. Workshop on Arabic and derived Script Analysis and Recognition (ASAR), 2018.*

Al-Ohali, Y., Cheriet, M. & Suen, C. (2003). Databases for recognition of handwritten Arabic cheques. *Pattern Recognition*, 36(1), 111 - 121.

Alaei, A., Raveaux, R. & Conte, D. (2017). Image quality assessment based on regions of interest. *Signal, Image and Video Processing*, 11(4), 673–680. Consulted at https://doi.org/10.1007/s11760-016-1009-z.

Alberti, M., Vögtlin, L., Pondenkandath, V., Seuret, M., Ingold, R. & Liwicki, M. (2019). Labeling, Cutting, Grouping: an Efficient Text Line Segmentation Method for Medieval Manuscripts. *CoRR*, abs/1906.11894.

Allier, B. & Emptoz, H. (2003, Sept). Character prototyping in document images using Gabor filters. *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, pp. I-537-40, vol.1.

Antonacopoulos, A. & Downton, A. C. (2007). Special Issue on the Analysis of Historical Documents. *Int. J. Doc. Anal. Recognit.*, 9(2), 75–77.

Arnia, F., Fardian, Muchallil, S. & Munadi, K. (2015, Aug). Noise characterization in ancient document images based on DCT coefficient distribution. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 971-975.

Asi, A., Cohen, R., Kedem, K., El-Sana, J. & Dinstein, I. (2014). A coarse-to-fine approach for layout analysis of ancient manuscripts. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 140–145.

Atena Shahkolaei,Azeddine Beghdadi, M. C. (2019). Blind Quality Assessment Metric and Degradation Classification for Degraded Document Images. *Signal Processing: Image Communication*, 76, 11-21.

Baechler, M. & Ingold, R. (2011, Sep.). Multi Resolution Layout Analysis of Medieval Manuscripts Using Dynamic MLP. *2011 International Conference on Document Analysis and Recognition*, pp. 1185-1189.

Basak, D., Pal, S. & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10), 203–224.

Ben Messaoud, I., Amiri, H., El Abed, H. & Margner, V. (2011, Sept). New Binarization Approach Based on Text Block Extraction. *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 1205-1209.

Blando, L., Kanai, J. & Nartker, T. (1995, 09). Prediction of OCR accuracy using simple image features. pp. 319-322 vol.1. doi: 10.1109/ICDAR.1995.599003.

Boukerroui, D., Noble, J. A. & Brady, M. (2004). On the choice of band-pass quadrature filters. *Journal of Mathematical Imaging and Vision*, 21(1-2), 53–80.

Bourgeois, F. L., Trinh, E., Allier, B., Eglin, V. & Emptoz, H. (2004). Document images analysis solutions for digital libraries. *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pp. 2-24.

Bukhari, S. S., Shafait, F. & Breuel, T. M. (2011, Sept). High Performance Layout Analysis of Arabic and Urdu Document Images. *2011 International Conference on Document Analysis and Recognition*, pp. 1275-1279.

Bukhari, S. S., Breuel, T. M., Asi, A. & El-Sana, J. (2012). Layout analysis for arabic historical document images using machine learning. *2012 International Conference on Frontiers in Handwriting Recognition (ICFHR 2012)*, pp. 639–644.

Cannon, M., Hochberg, J. & Kelly, P. (1999). Quality assessment and restoration of typewritten document images. *International Journal on Document Analysis and Recognition*, 2(2), 80–89.

Chandler, D. M. & Hemami, S. S. (2007). VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. *IEEE Transactions on Image Processing*, 16(9), 2284-2298.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3), 27.

Chetouani, A. & Beghdadi, A. (2011). Image quality assessment based on distortion identification. *Image Quality and System Performance VIII*, 7867, 78670E.

Chetouani, A., Beghdadi, A. & Deriche, M. (2012). A hybrid system for distortion classification and image quality evaluation. *Signal Processing: Image Communication*, 27(9), 948 - 960.

Chou, S. L. & Yu, S. S. (1993, Oct). Sorting qualities of handwritten Chinese characters for setting up a research database. *International Conference on Document Analysis and Recognition*, pp. 474-477.

Cohen, R., Asi, A., Kedem, K., El-Sana, J. & Dinstein, I. (2013). Robust text and drawing segmentation algorithm for historical documents. *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp. 110–117.

Daniel L Ruderman. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, 5(4), 517-548.

Eilertsen, G., Wanat, R., Mantiuk, R. K. & Unger, J. (2013). Evaluation of Tone Mapping Operators for HDR-Video. *Computer Graphics Forum*, 32(7), 275–284. Consulted at http://dx.doi.org/10.1111/cgf.12235.

Fadoua, D., Bourgeois, F. & Emptoz, H. (2006). Document Analysis Systems VII: 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006. Proceedings (ch. Restoring Ink Bleed-Through Degraded Document Images Using a Recursive Unsupervised Classification Technique, pp. 38–49). Berlin, Heidelberg: Springer Berlin Heidelberg. Consulted at http://dx.doi.org/10.1007/11669487_4.

Fang, Y., Yan, J., Li, L., Wu, J. & Lin, W. (2018). No Reference Quality Assessment for Screen Content Images With Both Local and Global Feature Representation. *IEEE Transactions on Image Processing*, 27(4), 1600-1610.

Farid, S. & Ahmed, F. (2009). Application of Niblack's method on images. *Emerging Technologies, 2009. ICET 2009. International Conference on*, pp. 280–286.

Ferwerda, J. A. (2008). Psychophysics 101: how to run perception experiments in computer graphics. *ACM SIGGRAPH 2008 classes*, pp. 87.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12), 2379–2394.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G. & Stolz, M. (2009, Sep.). Automatic Transcription of Handwritten Medieval Documents. *2009 15th International Conference on Virtual Systems and Multimedia*, pp. 137-142.

Gatos, B., Ntirogiannis, K. & Pratikakis, I. (2011). DIBCO 2009: document image binarization contest. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(1), 35-44. Consulted at http://dx.doi.org/10.1007/s10032-010-0115-7.

Geisler, W. (2008). Visual perception and the statistical properties of natural scenes, 2008. In *Annual Review of Psychology* (vol. 59, pp. 167–192).

Ghadiyaram, D. & Bovik, A. C. (2017). Perceptual Quality Prediction on Authentically Distorted Images Using a Bag of Features Approach. *Journal of Vision (in press)*.

Goltz, D., Attas, M., Young, G., Cloutis, E. & Bedynski, M. (2010). Assessing stains on historical documents using hyperspectral imaging. *Journal of Cultural Heritage*, 11(1), 19 - 26.

Govindaraju, V. & Srihari, S. N. (1995, Oct). Image quality and readability. *International Conference on Image Processing*, 3, 324-327, 1995.

Gu, K., Zhai, G., Yang, X. & Zhang, W. (2013, May). A new reduced-reference image quality assessment using structural degradation model. *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, pp. 1095-1098.

Gu, K., Zhai, G., Yang, X. & Zhang, W. (2015). Using Free Energy Principle For Blind Image Quality Assessment. *IEEE Transactions on Multimedia*, 17(1), 50-63.

Gu, K., Zhai, G., Lin, W., Yang, X. & Zhang, W. (2016). Learning a blind quality evaluation engine of screen content images. *Neurocomputing*, 196, 140 - 149.

Hasler, D. & Susstrunk, S. (2003). Measuring Colourfulness in Natural Images. *Proc. IS&T/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, 5007, 87–95.

Hedjam, R., Nafchi, H. Z., Kalacska, M. & Cheriet, M. (2015). Influence of Color-to-Gray Conversion on the Performance of Document Image Binarization: Toward a Novel Optimization Problem. *IEEE Transactions on Image Processing*, 24(11), 3637-3651.

Henriksson, L., Hyvarinen, A. & Vanni, S. (2009). Representation of Cross-Frequency Spatial Phase Relationships in Human Visual Cortex. *Journal of Neuroscience*, 29(45), 14342–14351. Consulted at http://www.jneurosci.org/content/29/45/14342.

Hripcsak, G. & Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3), 296–298.

I. Sobel, G. F. (1968). *A 3x3 isotropic gradient operator for image processing*. Presented at a talk at the Stanford Artificial Project.

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neurosciencevolume*, 2, 194-203.

Jain, A. K. & Farrokhnia, F. (1990, Nov). Unsupervised texture segmentation using Gabor filters. *1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings*, pp. 14-19.

Kang, L., Ye, P., Li, Y. & Doermann, D. (2014, Oct). A deep learning approach to document image quality assessment. *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 2570-2574.

Knox, K. T., Easton, R. L. & Christens-Barry, W. (2008, Aug). Image restoration of damaged or erased manuscripts. *2008 16th European Signal Processing Conference*, pp. 1-5.

Kovesi, P. (1999). Image Features From Phase Congruency. *Videre: J. Comput. Vis. Res.*, vol. 1, pp. 1-26.

Krasula, L. (2017). *Quality Assessment Methodologies of Post-Processed Images*. (Ph.D. thesis, Universite de Nantes).

Krzic, A. S., Donlic, M., Pejcinovic, M. & Sersic, D. (2016, May). Image sharpness assessment based on local phase coherence and LAD criterion. *2016 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 1-4.

Kulesh, V., Schaffer, K., Sethi, I. & Schwartz, M. (2001). Handwriting Quality Evaluation. In Singh, S., Murshed, N. & Kropatsch, W. (Eds.), *Advances in Pattern Recognition — ICAPR 2001: Second International Conference Rio de Janeiro, Brazil, March 11–14, 2001 Proceedings* (pp. 157–165). Berlin, Heidelberg: Springer Berlin Heidelberg. Consulted at http://dx.doi.org/10.1007/3-540-44732-6_16.

Kumar, D. & Ramakrishnan, A. G. (2011). QUAD: Quality Assessment of Documents. *4th Camera-based Document Analysis and Recognition (CBDAR 2011)*.

Kumar, J., Chen, F. & Doermann, D. (2012, Nov). Sharpness estimation for document and scene images. *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3292-3295.

Kumar., J., Ye., P. & Doermann., D. (2013). A Dataset for Quality Assessment of Camera Captured Document Images. *CBDAR. Lecture Notes in Computer Science, Springer, Cham*, 8357.

Kundu, D. & Evans, B. L. (2015). Full-reference visual quality assessment for synthetic images: A subjective study. *IEEE International Conference on Image Processing*, pp. 2374-2378.

Kundu, D., Ghadiyaram, D., Bovik, A. C. & Evans, B. L. (2017). No-Reference Quality Assessment of Tone-Mapped HDR Pictures. *IEEE Transactions on Image Processing*, 26(6), 2957-2971.

Lahouhou, A., Viennet, E. & Beghdadi, A. (2010). Selecting low-level features for image quality assessment by statistical methods. *Journal of computing and information technology*, vol. 18, 6.

Lasmar, N. E., Stitou, Y. & Berthoumieu, Y. (2009, Nov). Multiscale skewed heavy tailed model for texture analysis. *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2281-2284.

Lee, M.-S. & Medioni, G. (1999). Grouping into Regions, Curves, and Junctions. *Computer Vision and Image Understanding*, 76(1), 54-69.

Li, Q. & Wang, Z. (2009). Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation. *IEEE Journal of Selected Topics in Signal Processing*, 3(2), 202-211.

Liang, H. (2012). Advances in multispectral and hyperspectral imaging for archaeology and art conservation. *Applied Physics A*, 106(2), 309–323.

Lins, R. D. (2009). A Taxonomy for Noise in Images of Paper Documents - The Physical Noises. In Kamel, M. & Campilho, A. (Eds.), *Image Analysis and Recognition: 6th International Conference, ICIAR 2009, Halifax, Canada, July 6-8, 2009. Proceedings* (pp. 844–854). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lins, R. D., Banergee, S. & Thielo, M. (2010). Automatically Detecting and Classifying Noises in Document Images. *Proceedings of the 2010 ACM Symposium on Applied Computing*, (SAC '10), 33–39. Consulted at http://doi.acm.org/10.1145/1774088.1774096.

Liu, A., Lin, W. & Narwaria, M. (2012). Image Quality Assessment Based on Gradient Similarity. *IEEE Transactions on Image Processing*, 21(4), 1500-1512.

Liu, L., Dong, H., Huang, H. & Bovik, A. C. (2014). No-reference image quality assessment in curvelet domain. *Signal Processing: Image Communication*, 29(4), 494 - 505.

Liu, Z. & Laganiare, R. (2007). Phase congruence measurement for image similarity assessment. *Pattern Recognition Letters*, 28(1), 166-172.

Lu, H., Kot, A. C. & Shi, Y. Q. (2004). Distance-Reciprocal Distortion Measure for Binary Document Images. *IEEE Signal Processing Letters*, 11(2), 228-231.

Mantiuk, R. K., Tomaszewska, A. & Mantiuk, R. (2012). Comparison of four subjective methods for image quality assessment. *Computer graphics forum*, 31(8), 2478–2491.

Mehri, M., Gomez-Kramer, P., Heroux, P., Boucher, A. & Mullot, R. (2013). Texture feature evaluation for segmentation of historical document images. *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp. 102–109.

Min, X., Gu, K., Zhai, G., Liu, J., Yang, X. & Chen, C. W. (2018). Blind Quality Assessment Based on Pseudo Reference Image. *IEEE Transactions on Multimedia*, 1-1.

Mittal, A., Moorthy, A. K. & Bovik, A. C. (2012). No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12), 4695-4708.

Mittal, A., Soundararajan, R. & Bovik, A. (2013). Making a “Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3), 209-212.

Moghaddam, R. & Cheriet, M. (2010). A Variational Approach to Degraded Document Enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8), 1347-1361.

Montani, I., Sapin, E., Pahud, A. & Margot, P. (2012). Enhancement of writings on a damaged medieval manuscript using ultraviolet imaging. *Journal of Cultural Heritage*, 13(2), 226 - 228.

Moorthy, A. K. & Bovik, A. C. (2010). A Two-Step Framework for Constructing Blind Image Quality Indices. *IEEE Signal Processing Letters*, 17(5), 513-516.

Moorthy, A. K. & Bovik, A. C. (2011). Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing*, 20(12), 3350-3364.

Nafchi, H. Z., Moghaddam, R. F. & Cheriet, M. (2014). Phase-Based Binarization of Ancient Document Images: Model and Applications. *IEEE Transactions on Image Processing*, 23(7), 2916-2930.

Nafchi, H. Z., Shahkolaei, A., Moghaddam, R. F. & Cheriet, M. (2015). FSITM: A Feature Similarity Index For Tone-Mapped Images. *IEEE Signal Processing Letters*, 22(8), 1026-1029.

Nafchi, H. Z., Shahkolaei, A., Hedjam, R. & Cheriet, M. (2016). Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator. *IEEE Access*, 4, 5579-5590.

Obafemi-Ajayi, T. & Agam, G. (2012). Character-Based Automated Human Perception Quality Assessment in Document Images. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 42(3), 584-595.

Oppenheim, A. V. & Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, 69(5), 529-541.

Panichkriangkrai, C., Li, L. & Hachimura, K. (2013). Character segmentation and retrieval for learning support system of Japanese historical books. *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp. 118–122.

Papari, G. & Petkov, N. (2011). Edge and line oriented contour detection: State of the art. *Image and Vision Computing*, 29, 79 - 103.

Park, J.-S., Kang, H.-J. & Lee, S.-W. (2000). Automatic quality measurement of gray-scale handwriting based on extended average entropy. *15th International Conference on Pattern Recognition*, 4, 426-429 vol.4.

Peng, X., Cao, H., Subramanian, K., Prasad, R. & Natarajan, P. (2011, Sep.). Automated image quality assessment for camera-captured OCR. *2011 18th IEEE International Conference on Image Processing*, pp. 2621-2624.

Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F. & Kuo, C.-C. (2013, June). Color image database TID2013: Peculiarities and preliminary results. *4th European Workshop on Visual Information Processing (EUVIP)*, pp. 106-111.

Pottier, F., Michelin, A. & Robinet, L. (2019). Recovering illegible writings in fire-damaged medieval manuscripts through data treatment of UV-fluorescence photography. *Journal of Cultural Heritage*, 36, 183 - 190.

Pratikakis, I., Gatos, B. & Ntirogiannis, K. (2010). H-DIBCO 2010 - Handwritten Document Image Binarization Competition. *ICFHR'10*, 727-732.

Qureshi, M. A., Beghdadi, A. & Deriche, M. (2017). Towards the design of a consistent image contrast enhancement evaluation measure. *Signal Processing: Image Communication*, 58, 212 - 227.

Rehman, A. & Wang, Z. (2012). Reduced-Reference Image Quality Assessment by Structural Similarity Estimation. *IEEE Transactions on Image Processing*, 21(8), 3378-3389.

Rowley-Brooke, R. & Kokaram, A. (2012). Bleed-through removal in degraded documents. Consulted at http://dx.doi.org/10.1117/12.908911.

Saad, M. A., Bovik, A. C. & Charrier, C. (2012). Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. *IEEE Transactions on Image Processing*, 21(8), 3339-3352.

Saha, A. & Wu, Q. J. (2013). Perceptual image quality assessment using phase deviation sensitive energy features. *Signal Processing*, 93(11), 3182-3191.

Savino, P. & Tonazzini, A. (2016). Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs. *Journal of Cultural Heritage*, 19, 511-521.

Shahkolaei, A., Ziaei Nafchi, H., Al-Maadeed, S. & Cheriet, M. (November 2017). Subjective and Objective Quality Assessment of Degraded Document Images. *Journal of Cultural Heritage*, no. 11.

Sharifi, K. & Leon-Garcia, A. (1995). Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(1), 52-56.

Sharma, G. (2000, Sept). Cancellation of show-through in duplex scanning. *International Conference on Image Processing*, 2, 609-612 vol.2.

Sheikh, H. & Bovik, A. (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2), 430-444.

Sheikh, H., Sabir, M. & Bovik, A. (2006). A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *Image Processing, IEEE Transactions on*, 15(11), 3440-3451.

Sun, W., Zhou, F. & Liao, Q. (2017). MDID: A multiply distorted image database for image quality assessment. *Pattern Recognition*, 61, 153-168.

United Nations Educational, Scientific and Cultural Organization (UNESCO).

van de Weijer, J., Gevers, T. & Bagdanov, A. D. (2006). Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1), 150-156.

Van Phan, T., Zhu, B. & Nakagawa, M. (2011). Development of Nom character segmentation for collecting patterns from historical document pages. *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pp. 133–139.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer Verlag.

Villegas, M. & Toselli, A. H. (2014, Sept). Bleed-Through Removal by Learning a Discriminative Color Channel. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 47-52.

Vu, C., Phan, T., Singh, P., & Chandler, D. M. (2012). Digitally Retouched Image Quality (DRIQ) Database.

Walvoord, D. J. & Easton, R. L. (2008). Digital Transcription of the Archimedes Palimpsest [Applications Corner]. *IEEE Signal Processing Magazine*, 25(4), 100-104.

Wang, Z. & Simoncelli, E. (2004). Local phase coherence and the perception of blur. In *Advances in Neural Information Processing Systems 16 - Proceedings of the 2003 Conference, NIPS 2003*. Neural information processing systems foundation, 2004, Vol. 16.

Wang, Z. & Simoncelli, E. P. (2005). Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. *Proc. SPIE, Conf. on Human Vision and Electronic Imaging, X*, 5666, 149-159. Consulted at http://dx.doi.org/10.1117/12.597306.

Wang, Z., Sheikh, H. R. & Bovik, A. (2002). No-reference perceptual quality assessment of JPEG compressed images. *International Conference on Image Processing. 2002*, 1, I-477-I-480 vol.1.

Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.

Wang, Z., Wu, G., Sheikh, H. R., Simoncelli, E. P., Yang, E.-H. & Bovik, A. C. (2006). Quality-aware images. *IEEE Transactions on Image Processing*, 15(6), 1680-1689.

Williams, L. R. & Thornber, K. K. (1999). A comparison of measures for detecting natural shapes in cluttered backgrounds. *International Journal of Computer Vision*, 34(2-3), 81–96.

Winkler, S. (2012). Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6), 616–625.

Xie, Y., Lu, H. & Yang, M.-H. (2013). Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing*, 22(5), 1689–1698.

Xu, J., Ye, P., Li, Q., Liu, Y. & Doermann, D. (2016, Sept). No-reference document image quality assessment based on high order image statistics. *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3289-3293.

Xue, W., Zhang, L., Mou, X. & Bovik, A. C. (2014). Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Transactions on Image Processing*, 23(2), 684-695.

Yang, H., Fang, Y. & Lin, W. (2015). Perceptual Quality Assessment of Screen Content Images. *IEEE Transactions on Image Processing*, 24(11), 4408-4421.

Ye, P. & Doermann, D. (2012, Nov). Learning features for predicting OCR accuracy. *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3204-3207.

Ye, P., Kumar, J., Kang, L. & Doermann, D. (2013, June). Real-Time No-Reference Image Quality Assessment Based on Filter Learning. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 987-994.

Zhang, L., Zhang, D., Mou, X. & Zhang, D. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(8), 2378-2386.

Zhang, Y. & Chandler, D. (2013). An Algorithm for No-Reference Image Quality Assessment Based on Log-Derivative Statistics of Natural Scenes. *Journal of Electronic Imaging*, 22. doi: 10.1117/12.2001342.

Zhou, J., Xiao, B. & Li, Q. (2008, June). A no reference image quality assessment method for JPEG2000. *IEEE International Joint Conference on Neural Networks*, pp. 863-868.