Screen content image evaluation and processing

Yang, Huan

2015

Yang, H. (2015). Screen content image evaluation and processing. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/65397>

<https://doi.org/10.32657/10356/65397>

SCREEN CONTENT IMAGE EVALUATION AND PROCESSING

YANG HUAN

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy (Ph.D)

2015

To my family, for their unconditional love and endless support.

Acknowledgement

First and foremost I offer my heartfelt thanks to my supervisor, Prof. Weisi Lin, for his support, guidance, encouragement, and patience. It has been a great pleasure to work with him, whose personality inspired me a lot as well as his knowledge did.

Special thanks to my collaborators, Asst. Prof. Yuming Fang, Prof. Shiqian Wu and Prof. Zhou Wang for their invaluable suggestions and comments.

I would also like to express my sincere thanks to my labmates and friends, for their lovely company and encouragement. I always appreciate the precious moments we shared together. I also thank my colleagues and technicians in CeMNet for their patient and kind help.

Finally, My deepest gratitude goes to my family for their continuous love and support throughout my life. I owe to them every step of my progress. I must thank my husband, Feng Li. He was always there cheering me up and stood by me through the good times and bad.

Abstract

Screen Content Image (SCI) is a typical kind of compound images that contain texts, graphics and pictures concurrently. With the rapid development of digital devices and computing techniques, SCIs have increasingly appeared in multi-client communication systems. The related applications bring many challenges on SCI processing, such as acquisition, segmentation, compression, transmission, quality evaluation, etc. SCIs have different characteristics from natural scene images and scanned document images, which result in the fact that existing classical image processing methods cannot effectively process SCIs. Hence, specialized algorithms for SCI processing are much desired. Currently, there is no much research work in the literature for SCI processing. In this research work, we try to understand the basic properties of SCIs and focus on addressing challenging problems in the following three aspects, i.e., segmentation, compression and perceptual quality assessment of SCIs.

SCI segmentation, which aims to distinguish texts from other components, is a fundamental step in various SCI processing techniques. In this research work, we firstly propose a coarse-to-fine framework to segment texts with arbitrary scales and orientations from other components in SCIs. A *Local Image Activity Measure* (LIAM) is designed to enhance the difference between textual and pictorial regions and eliminate most of pictorial regions with low frequency. In order to remove survived pictorial regions (mistaken as texts), a new *Scale and Orientation Invariant Grouping* (SOIG) algorithm is proposed to construct *Textual Connected Components* (TCCs) with uniform geometrical features. False positive components are finally filtered out by three verification criteria. The proposed text segmentation algorithm can maintain integrity of texts with varied scales and orientations, which benefits the compression and evaluation procedures for SCIs.

It has been demonstrated that traditional coding methods with a single basic function, such as JPEG and JPEG2000, cannot achieve good performance for SCI compression due to the intensive high frequency variations in textual regions. In this work, a novel SCI compression scheme is proposed by using different basis functions to encode different components respectively. A tailored text dictionary for textual image representation is learned via a modified dictionary learning method, i.e., *K-Singular Value Decomposition* (K-SVD). Compared with the *Discrete Cosine Transform* (DCT) based representation, textual representation derived from the tailored text dictionary is much sparser, which provides more probability to effectively encode SCIs. The proposed coding scheme achieves much higher coding performance than existing standard coding methods, especially for SCIs with large percentage of textual regions.

To evaluate the visual quality of the processed SCIs by compression and other processing, we present a study on perceptual quality assessment of SCIs. A large *SCI Quality Assessment Database* (SIQAD) is constructed with the visual quality scores obtained through subjective testing. Besides, we investigate the correlations between the subjective scores of different regions, which reveals the impact of textual and pictorial regions to the overall visual quality. A new *SCI Perceptual Quality Assessment* (SPQA) scheme is also proposed to automatically evaluate the visual quality of distorted SCIs, by taking into account the different properties and contributions of textual and pictorial regions. Compared with the start-of-the-art *Image Quality Assessment* (IQA) methods, the proposed SPQA achieves much higher consistency with subjective results.

Keywords

Screen content images, text segmentation, textual image coding, subjective quality testing, objective quality metrics, screen content compression.

Contents

Acknowledgement	iii
Abstract	v
List of Figures	xi
List of Tables	xv
List of Abbreviations	xvii
1 Introduction	1
1.1 Background	2
1.2 Scope of Research	5
1.3 Contributions	7
1.4 Thesis Organization	9
2 Literature Review	11
2.1 Properties of Screen Content Images	11
2.2 Text Segmentation for Screen Content Images	14
2.3 Coding Algorithms for Screen Content Images	16
2.4 Perceptual Quality Assessment of Screen Content Images	19
3 Scale and Orientation Invariant Text Segmentation	27
3.1 Introduction	28
3.2 Local image activity measure (LIAM) based coarse partition	30
3.3 Textual connected component (TCC) based refinement	34
3.3.1 Extraction of TCCs with Minimum Scale	35
3.3.2 Scale and Orientation Invariant Grouping (SOIG)	36

3.3.3	Criteria for TCC Verification	42
3.4	Experimental Results	45
3.4.1	Effectiveness Analysis of the proposed LIAM	45
3.4.2	Text Segmentation On ICDAR Database	48
3.4.3	Text Segmentation on the BCID database	51
3.4.4	Impact of Segmentation Accuracy on SCI Compression	55
3.4.5	Further Analysis of the Proposed Algorithms	58
3.5	Summary	61
4	Learning Based Compression for Screen Content Images	63
4.1	Introduction	63
4.2	Over-complete DCT based Text Dictionary Learning	64
4.2.1	Over-complete DCT based K-SVD Algorithm	65
4.2.2	Text Dictionary Learning	67
4.2.3	Sparsity Analysis of Text Representation	69
4.3	The Proposed SCI Compression Scheme	71
4.3.1	Preprocessing of SCIs	71
4.3.2	Coding of Sparse Textual Representation Matrix	72
4.4	Experimental Results	73
4.5	Summary	77
5	Perceptual Quality Assessment of Screen Content Images	79
5.1	Introduction	79
5.2	Construction of the Screen Image Quality Assessment Database (SIQAD)	81
5.2.1	Generation of Distorted SCIs in the <i>SIQAD</i>	81
5.2.2	Equipment Configuration and Human Subjects	83
5.2.3	Subjective Testing Methodology: Absolute Category Rating (ACR) for Assessing Quality of SCIs	85
5.3	Perceptual Quality Assessment Of SCIs	86
5.3.1	Which Part Attracts More Visual Attention: Textual or Pictorial?	87
5.3.2	Objective Quality Assessment of SCIs	89
5.4	Experimental Results	93
5.4.1	Reliability of DMOS	93
5.4.2	Analysis of Subjective Scores of Different Regions	95

5.4.3	Validation of the Proposed SPQA	96
5.4.4	Proposed Weighting Strategy on NIQA Methods	102
5.5	Summary	103
6	Conclusion	105
6.1	Research Contributions	105
6.2	Future Directions	106
	Bibliography	109
	Publications	121

CONTENTS

List of Figures

1.1	Examples of SCIs with various content styles. (a) and (b) are cut from webpages; (c) is an email attached poster; (d) is from digital magazine; (e) is an advertisement poster; (f) is a slide page.	2
1.2	The platform of the virtualized screen sharing system [1]	3
1.3	The diagram of the deep shot system [2] that transmits application states from computer to mobile phones.	4
2.1	Some examples of natural scene images and scanned document images. In (a) and (b), there are two natural images with texts mixed in the scenes. In (c) and (d), two scanned document images are shown with various noise artifacts.	12
2.2	Comparison of some texts on the three kinds of images. In (b) and (c), text strings are amplified to show the detail information.	15
2.3	Illustration of MRC mode for document compression. The foreground and background layers are encoded through standard encoding algorithms (e.g., JPEG and JPEG2000), and the binary mask is encoded by using JBIG or JBIG2.	17
2.4	Examples of natural images, textual images and screen content images.	23
2.5	Distribution of values in the naturalness maps of the example images. .	24
2.6	Local activity value distributions for natural and textual images.	25
3.1	The proposed coarse-to-fine text segmentation framework. Three main contributions of the proposed method are highlighted by filling.	29
3.2	Four activity measures: D1-HV, D1-Diag, D2-HV and D2-Diag.	31
3.3	Average activity values of tested images using different measures.	31

3.4	LIAM activity map and its binarization result for one SCI.	33
3.5	Flowchart of the proposed TCC-based refinement, including the three key steps highlighted by color filling.	35
3.6	Illustration of intra-string and inter-string distance between text characters.	37
3.7	Calculation of the DAD between two point sets: A and B . $d(a_i, b_j)$ computes the nearest Euclidean distance from A to b_j	38
3.8	The flowchart of the SOIG algorithm. The neighbour searching is implemented based upon the defined three rules. The adaptive computation of θ and L_{mean} is given in detail in Algorithm 1.	40
3.9	An orientation-related structure kernel. Four central lines are marked by dotted lines and the direction of these lines is 30°	41
3.10	Results of intermediate steps in the TCC-based refinement.	42
3.11	Feature analysis of different candidate connected components.	43
3.12	Comparison of binarized results of activity maps generated by different methods.	47
3.13	Computation of pixel variation in (a) SFM and (b) IAMx.	48
3.14	Experiment results for some samples in ICDAR 2011. In (b) and (c), the left is the ground truth and the right is our result.	50
3.15	Comparison of text detection results. The left images are results using the method in [3], and the right ones are results of our method.	51
3.16	Segmentation results of example I by the four methods. The segmented textual layers are displayed in (b)-(e), respectively (the padded pixels are set to 0 in the textual layers).	53
3.17	Segmentation results of example II by the four methods. The segmented textual layers are displayed in (b)-(e), respectively (the padded pixels are set to 0 in the textual layers).	54
3.18	Text detection of [4] of a tested image (the padded pixels are set to 0 in the segmented layers).	55
3.19	Visual quality of reconstructed image based on different coding methods. The bitrates for all these images are 0.4 bpp. (f) is based on the proposed segmentation algorithm.	58
3.20	Compression results for SCIs.	59

3.21	Connection results using different distance measure.	60
3.22	Text detection results for artistic images using our algorithm.	62
4.1	The proposed scheme for SCIs compression. The key steps in the scheme are highlighted by color filling in the figure.	65
4.2	Over-complete $DCT441$ (left) and D^T441 (right).	68
4.3	Reconstruction results based on different dictionaries.	69
4.4	Energy compaction of non-zero coefficients.	70
4.5	Position-based coding for the sparse coefficient matrix C'	73
4.6	Compression results for four example SCIs: (a) is a webpage with binary texts; (b) is a webpage with colored texts; a slide image and a combined screen image are shown in (c) and (d).	74
4.7	Quality comparison of reconstructed images (bitrate:1.0 bpp).	75
4.8	RD curves of Compression results for four screen images.	76
5.1	Examples of distorted SCIs. Due to the space limitation, only one level distortion is illustrated (refer to the database for more images).	84
5.2	Graphical user interface in the subjective test. The red tooltip will change if subjects need to judge different regions.	85
5.3	Diagram of the proposed algorithm SPQA. The SPQA mainly contains two algorithms highlighted in the figure.	87
5.4	DMOS values of some examples in <i>SIQAD</i> . The scale of the DMOS values is from 0 to 100. A higher value represents worse visual quality of the image (refer to the images at the original resolution for better visual comparison).	90
5.5	Filters for calculating the sharpness values.	92
5.6	Distributions of DMOS values of two examples. The error bars indicate the confidence intervals of related scores.	94
5.7	Histogram of DMOS values of images in the <i>SIQAD</i>	95
5.8	Scatter of predicted quality scores by some metrics against the DMOS values on the <i>SIQAD</i> . The vertical axis in each figure is the DMOS values.	98
5.9	Visual quality comparison of SCIs with different distortion types. The DMOS values and the quality scores predicted by four different metrics (PSNR, SSIM, VIF and SPQA) are provided for comparison.	101

LIST OF FIGURES

List of Tables

3.1	Comparison of text detection results on artificial SCIs.	48
3.2	Experiment results on ICDAR database.	49
3.3	Experimental results on BCID database	56
4.1	Convergence Comparison Result of overcomplete DCT based Learning and random sample based learning.	67
4.2	PSNR values (dB) for test images at given bitrates.	72
5.1	Correlation analysis of the obtained quality scores for the entire images, textual and pictorial regions.	96
5.2	Comparison of two combination methods	97
5.3	Correlation results (PLCC) of the DMOS values and the objective scores given by 12 metrics. The paired T-test is applied to the proposed SPQA against the 11 NIQA methods. The result ($H = 1, P < 0.05$) for each pair indicates that the SPQA is significantly better than the tested 11 NIQA methods.	99
5.4	Correlation results (SROCC) of the DMOS values and the objective scores given by 12 metrics.	100
5.5	Correlation results (RMSE) of the DMOS values and the objective scores given by 12 metrics.	100
5.6	Comparison of two combination methods	103

LIST OF TABLES

List of Abbreviations

SCI	Screen Content Image
HVS	Human Visual System
IQA	Image Quality Assessment
PSNR	Peak Signal Noise Ratio
JPEG	Joint Photographic Experts Group
QoE	Quality of Experience
LIAM	Local Image Activity Measure
TCC	Textual Connected Component
SOIG	Scale and Orientation Invariant Grouping
MAD	Minimum Average Distance
DCT	Discrete Cosine Transformation
SIQAD	Screen Image Quality Assessment Database
ACR	Absolute Category Rating
SPQA	SCI Perceptual Quality Assessment
K-SVD	K-Singular Value Decomposition
MRC	Mixed Raster Content
NIQA	Natural Image Quality Assessment
DIQA	Document Image Quality Assessment
SIQA	Screen Image Quality Assessment
DMOS	Difference of Mean Opinion Scores
LSC	Layer Segmentation based Coding
PLCC	Pearson Linear Correlation Coefficient
SROCC	Spearman Rank-Order Correlation Ccoefficient
RMSE	Root Mean Squared Error

Chapter 1

Introduction

Screen Content Images (SCIs), which render texts, graphics and pictures together on digital devices, have been attracting great interests in both academic and industrial communities, given their pervasive applications in many fields, e.g., virtualized screen sharing systems [1], multi-client communication systems [2], remote computing platforms [5, 6, 7], remote education, product advertising, etc. In these applications, visual content (e.g., web pages, emails, slide files and computer screens) is typically rendered in the form of SCIs, and then transmitted between different digital devices (computers, tablets or smart phones). Some examples of SCIs are provided in Fig. 1.1.

In the above mentioned SCI based applications, a lot of processing operations for SCIs are involved, including image acquisition, compression, transmission, reconstruction, storage, quality evaluation, etc. Among these processes, SCI compression is most oft-encountered. Although numerous coding algorithms have been proposed for natural scene image compression, such as JPEG, JPEG2000 and H.264 intra coding, the poor effectiveness of them on coding SCIs has been demonstrated [8, 9]. Hence, specialized compression methods for SCI compression with high efficiency is much desired. Currently, several studies have started to investigate segmentation based coding methods, in which texts and pictures are separately encoded [10, 11, 12, 13]. It is also important to build standard evaluation metrics to assess the visual quality of compressed SCIs with high consistency with the *Human Vision System* (HVS). In this thesis, we investigate segmentation, compression and quality evaluation of SCIs, and aim at designing novel algorithms to address challenging problems in these three aspects.

Figure 1.1: Examples of SCIs with various content styles. (a) and (b) are cut from web-pages; (c) is an email attached poster; (d) is from digital magazine; (e) is an advertisement poster; (f) is a slide page.

In this chapter, we first introduce the background and motivation behind this thesis to drive our research work in Sec. 1.1, before presenting the scope of this thesis in designing specific processing methods of SCIs in Sec. 1.2. We then summarize our contributions and outline this thesis in Secs. 1.3 and 1.4 respectively.

1.1 Background

SCIs have been involved in various applications. For example, in the virtualized screen sharing system shown in Fig. 1.2, the screen rendering is operated in the cloud and the rendered screens (i.e., SCIs) are delivered to multi-clients as part of the cloud services. In contrary to existing screen sharing systems (e.g., Virtual Network Computing (VNC) [14] and Remote Desktop Protocol (RDP) [15]) that render screens on the client devices, the virtualized screen sharing system has many advantages, such as providing

computationally intensive and graphically rich services to thin-client mobile devices, and transmitting images without considering the format conversion between different platforms [1]. Two cloud-mobile communication systems, Cloud Browser and Cloud Phone, have been reported in [1] to verify the advantages of such virtualized screen sharing mechanism. In the multi-client communication system [2], the deep shot soft-

Figure 1.2: The platform of the virtualized screen sharing system [1]

ware is firstly installed on all the client devices (computers and mobile phones). As shown in Fig. 1.3, an user takes a computer screen picture in a lab environment using his/her mobile phone. The deep shot then compares the captured picture with screen images of all registered computers in the lab, and finds the one displaying the picture. The deep-shot system can find the application which the user is looking through on the computer, and automatically migrates the application to the mobile phone and resumes its current state. With the popularization of social networking services, such as facebook, twitter and microblog, more and more contents (texts and pictures) are generated in the form of SCIs and uploaded to the Internet. These SCIs are collected and used in many text information retrieve systems or some personal behavior analysis applications. SCIs are also involved in many other applications, such as remote education systems, advertisement poster design, digital magazines, and so on.

Figure 1.3: The diagram of the deep shot system [2] that transmits application states from computer to mobile phones.

As aforementioned, many processing techniques are much desired in such applications, including SCI segmentation, compression, transmission and storage. Many algorithms have been reported to process screen content images/videos, especially text segmentation and compression of SCIs. It has been demonstrated that the general encoding methods, such as JPEG, JPEG2000 and H.264 coding algorithms, cannot encode screen content images and videos efficiently due to the specific characteristics of texts [8], since it is hard for them to efficiently encode the high frequency information in textual regions. Therefore, some text-segmentation based coding methods are reported, in which SCIs are segmented into different parts and all the parts are encoded separately by using different algorithms [8, 9, 10, 11, 12, 13, 16, 17]. These text-segmentation based coding methods achieve higher coding performance than the general coding methods. However, the performance of neither the text segmentation or the coding algorithm can be further improved. Due to the importance of SCI compression, MPEG/VCEG called for proposals to efficiently compress screen content images and videos with high visual consistency with the HVS as an extension of the HEVC standard [18], and currently, there are several proposals have been reported for screen

content coding [19, 20, 21, 22, 23, 24].

To evaluate the performance of coding methods, the visual quality assessment of processed images is also indispensable in the whole processing framework. When processing SCIs, various distortions may be involved, such as noise, blurring, contrast change and compression artifacts. For example, when SCIs are captured by smart phones, blurring appears on images along with hand-shake or out-of-focus of the camera. Different settings of brightness or contrast of screens will result in the contrast change of captured SCIs. Compression artifacts (e.g. blocking and quantization noises) commonly appear on compressed SCIs. Generally, *Peak Signal-to-Noise Ratio* (PSNR) is adopted in the aforementioned proposals to evaluate the quality of processed SCIs. However, it is known that PSNR is not consistent with human visual perception [25, 26]. In the competition organized by MPEG/VCEG [18], the coding results are also subjectively compared by the examiners. However, subjective testing is time and effort consuming, and not suitable for online and in-service operations. Hence, it is meaningful to build objective metrics to automatically assess the visual quality of SCIs.

Up to now, *Quality of Experience* (QoE) has been investigated to evaluate users' viewing experience on webpages, which is called Web QoE [27]. Unfortunately, the current Web QoE mainly focuses on some *Quality of Service* (QoS) metrics, e.g. loss ratio, full rendering and round-trip time, rather than taking differences of human perception for pictures and texts into account [28, 29]. In this case, if the overall loss ratio is determined, the QoS value would be constant. However, different loss ratios to pictorial and textual parts may lead to quite different QoE. Therefore, perceptual quality assessment of SCIs is much desired for various applications. Although many IQA methods have been proposed to evaluate quality of natural images [30], whether these IQA methods can be applicable to SCIs is still an open question.

1.2 Scope of Research

In this thesis, we try to investigate the differences among SCIs, natural scene images and scanned document images, and then build specialized processing methods for SCIs. In particular, considering the properties of SCIs, we propose novel algorithms to address challenging problems in segmentation, compression and quality assessment of SCIs. In

the SCI related applications, these three processing techniques are closely linked and interact with each other.

We firstly focus on text segmentation in SCIs, since it is a fundamental step in SCI compression and other processing applications (e.g., retrieving text information from SCIs). Most of advanced SCI coding methods are based on text segmentation, as textual regions have quite different statistical features from pictorial regions, and the text misclassification will degrade the final coding performance greatly. We propose a coarse-to-fine framework for segmenting texts with arbitrary scales and orientations in SCIs. In the coarse stage, the *Local Image Activity Measure* (LIAM) is designed based upon the variation distribution of characters, to highlight the difference between textual and pictorial regions. This stage outputs a coarse textual layer including textual regions as well as a few pictorial regions with high activity. In the fine stage, a Textual Connected Component(TCC) based refinement is proposed to eliminate the survived pictorial regions. In particular, a Scale and Orientation Invariant Grouping (SOIG) algorithm is proposed to adaptively generate TCCs with uniform statistical features. The Minimum Average Distance (MAD) and morphological operations are employed to assist the formation of candidate TCCs. Then, three string-level features (i.e. shapeness, color similarity and mean activity level) are designed to distinguish the true TCCs from the false positive ones that are formed by connecting the high activity pictorial components. The proposed text segmentation algorithm can be utilized to support the compression and quality evaluation procedures.

Based on the proposed text segmentation algorithm, we further design a novel encoding algorithm for SCIs. Generally, the SCIs can be mainly classified into two parts: textual and pictorial regions. The pictorial part can be encoded efficiently by traditional standard coding methods (e.g., JPEG, JPEG2000), while the effectiveness of these approaches to compress textual regions is still far away from being satisfactory due to the intensive high-frequency variations in the textual regions. In the proposed method, a tailored dictionary is constructed for text representation via a learning approach. Compared with the *Discrete Cosine Transformation* (DCT) based representation, the textual representation derived from the learned dictionary is much sparser. A novel SCI compression scheme is then proposed, employing two different basis functions to represent the textual and pictorial regions. The pictorial regions are encoded through DCT transform and entropy coding, while the textual regions are represented based on

the learned text dictionary. A position based coding strategy is adopted to assist the coding of the sparse textual coefficients.

When compressing SCIs, the visual quality of compressed images should be ensured, since these images are finally presented to human beings. Hence, visual quality assessment methods should be utilized to compare the final coding performance, rather than the commonly used PSNR. To evaluate the visual quality of SCIs, we present a study on perceptual quality assessment of distorted SCIs subjectively and objectively. We construct a large-scale Screen Image Quality Assessment Database (SIQAD) consisting of 20 source and 980 distorted SCIs. In order to get the subjective quality scores and investigate which part (text or picture) contributes more to the overall visual quality, three subjective scores corresponding to the entire, textual and pictorial regions are rated by subjects. The 11-category *Absolute Category Rating* (ACR) is employed to obtain the subjective data due to its practicability and flexibility. Based on the subjective scores, we propose a weighting model to account for the correlation between these three subjective scores. Furthermore, we design an objective metric to measure the quality of distorted SCIs, considering the visual difference between textual and pictorial regions. Experimental results demonstrate that the proposed *SCI Perceptual Quality Assessment* (SPQA) scheme, consisting of the objective metric and the weighting model, can obtain better performance than 11 state-of-the-art Image Quality Assessment (IQA) methods. To the best of our knowledge, the SIQAD is the first large-scale database published for quality evaluation of SCIs, and this is the first attempt to explore the perceptual quality assessment of distorted SCIs.

1.3 Contributions

In this thesis, we have made the following contributions:

- A coarse-to-fine text segmentation framework is proposed to segment texts in SCIs.
 - The proposed framework is robust to detect texts with arbitrary scales and orientations in SCIs with clutter background, which benefits from the proposed SOIG algorithm and the string-level feature extraction and identification.

- The proposed algorithm can preserve the integrality of text characters, avoiding over- or under-segmentation of textual regions, to favor the processing that follows (e.g., to maintain the readability of characters when retrieving texts or ensure the visual quality of textual regions when compressing SCIs).
- For characters embedded or superimposed on complex pictorial regions, over-connection can be well addressed to reduce the misclassification of text characters and pictorial regions.
- We propose a learning based compression method for SCIs, in which pictorial regions are encoded by using traditional DCT and textual regions are encoded by the learned dictionary.
 - The over-complete K-SVD algorithm is designed to learn the textual dictionary, which can achieve more time-efficient convergence than the random sample based K-SVD algorithm.
 - The representation of textual regions by using the learned dictionary is much sparser than that by using DCT. Cooperating with the position-based coding manner, the proposed learning based coding method achieves better performance than the standard image coding methods (e.g., JPEG and JPEG2000).
- We explore the perceptual quality assessment of distorted SCIs from both subjective and objective aspects.
 - We build a large-scale SCI quality assessment database (SIQAD), in which the DMOS values of all distorted images are provided.
 - Three subjective scores are obtained for each image in the database, according to textual, pictorial and entire regions, respectively. Through the analysis of the subjective data, we find that textual regions generally attract more attention than pictorial regions when viewing SCIs.
 - We propose a novel objective metric to assess visual quality of distorted SCIs, and demonstrate that the proposed metric can obtain higher consistency with the HVS than existing IQA methods.

1.4 Thesis Organization

The rest of this thesis is organized as follows.

In Chapter 2, we survey the literature related to the topics covered by this thesis, including text segmentation, compression and visual quality assessment of SCIs. We exploit the motivations to drive our research through carefully reviewing the representative papers. We also give a brief discussion upon the differences between our research threads and the existing works to further highlight our contributions.

In Chapter 3, we present our original work on text segmentation of SCIs. We present a new scale and orientation invariant algorithm for segmenting texts from SCIs. By using the proposed algorithm, texts can be segmented out exactly.

In Chapter 4, we propose a learning based sparse coding method for SCI compression. In our method, textual regions are encoded by the learned textual dictionary, such that the representation of textual regions is much sparser and the final coding performance is improved.

In Chapter 5, we study the perceptual quality measure of distorted SCIs with both subjective and objective testing. A large-scale image database is built with subjective scores. The proposed objective metric is then built based on the correlation analysis of these subjective scores.

In Chapter 6, we conclude the work we have done. We also outline some promising direction by extending our current research work.

Chapter 2

Literature Review

As aforementioned, SCIs are compound images that generated on digital devices by rendering various components into images and include textual and pictorial regions. There are other two kinds of compound images: natural scene images with texts and scanned document images. In this chapter, we first investigate the properties of SCIs by comparing SCIs with natural scene text images and scanned document images. We then introduce the existing work for processing SCIs, including text segmentation, compression and visual quality assessment. Text segmentation is a fundamental step for advanced segmentation-based SCI compression and quality assessment methods. The visual quality assessment methods for SCIs can be utilized in turn to guide the compression procedure to achieve higher compression performance. We also study the state-of-the-art methods on these three topics for natural scene images with texts and scanned document images.

2.1 Properties of Screen Content Images

In this section, we will investigate the properties of SCIs. Some examples of other two kinds of images, i.e., natural scene images and scanned document images, are given in Fig. 2.1 for viewing comparison. SCIs are different from natural scene images with texts and scanned document images from many aspects [8, 31]:

- Generation

SCIs are generated on digital devices by rendering various components into images. They can be captured by screen shot softwares on digital devices. Generally,

(a)

(b)

(c)

(d)

Figure 2.1: Some examples of natural scene images and scanned document images. In (a) and (b), there are two natural images with texts mixed in the scenes. In (c) and (d), two scanned document images are shown with various noise artifacts.

the resolutions of SCIs are quite low (usually less than 100dpi), depending on the configuration of digital devices. Natural scene images are generally captured by digital cameras with lens. The resolution of natural images are usually quite high due to the rapid development of digital cameras. Document images are generally scanned to capture paper-based information and hence, they usually have much higher resolution (larger than 300dpi) than SCIs.

- Composition

SCIs are composed of texts and graphics/pictures in various composition manners: for example, texts are gathered on smooth background, enclosed by pictures or superimposed on pictures. In natural scene images, texts are generally mixed with pictures and have no obvious color transition in edge regions, while in scanned document images, pictures are generally surrounded by binary or colored texts. Besides, due to the different generation mechanisms, the geometric features of texts and the ways to construct the basic textual objects for these three different kinds of images are significantly different. We will introduce more about the difference in the following sections.

- Noise affection

Since SCIs are automatically rendered on digital devices, less external noise appear on SCIs except the intrinsic distortions in pictures. When capturing natural images with texts, effect of lights, object moving or occlusion may occur, which lead to the difference on processing natural images with texts from SCIs. In scanned document images, as shown in Fig. 2.1 (c) and (d), a lot of noises appear, such as inking, blurring, soiling, color fading and human scribble. Hence, many processing methods for scanned document image focus on eliminating such pollution.

- Processing requirement

When processing scanned document images, taking compression as example, distortion of textual regions in document images can be somehow tolerated, since these images inevitably contain noise when generated. That is why sub-sampling is commonly used in compression of document images. However, the artifacts in textual regions in SCIs are easily noticeable. Therefore, the accuracy of SCI

segmentation should be high and more bits should be assigned to textual regions to ensure the visual quality of encoded SCIs.

2.2 Text Segmentation for Screen Content Images

Detecting texts from natural scene text images and scanned document images has been intensively studied during the past decades. However, text detection for SCIs is emerging recently, and is different from detecting text for the first two scenarios as explained next.

As aforementioned, natural scene images with texts are captured by cameras with high resolutions and only a few texts are mixed in the scenes, as shown in Fig. 2.1 (a) and (b). In such kind of images, the most prominent characteristic of texts is that stroke width of a character is large. Meanwhile, characters in one text string are separately distributed with similar color and stroke width, as shown in the left-hand side of Fig. 2.2 (a). Hence, stroke width, color uniformity and other character-level features (e.g. occupy-rate, aspect-ratio and compactness) can be extracted to identify candidate characters on natural scene text images [3, 32, 33, 34, 35, 36]. However, texts in SCIs are usually embedded on complex background with arbitrary colors, scales and orientations in low resolutions. Color transition of character edges is obvious and strokes of characters are generally slight. Moreover, some characters are connected together, as shown in the right-hand side of Fig. 2.2 (a). Hence, using the above methods to segment texts in SCIs, it is difficult to extract uniform color or stroke width features from small characters and to ensure the uniformity of character-level features for connected characters, such that small or connected characters are easily ignored. We will further demonstrate this in our experiments in Sec. 3.4.2.

Many methods have been proposed for segmenting texts from scanned document images [4, 37, 38, 39, 40, 41, 42, 43, 44]. Traditionally, the research works mainly focus on the scanned document images with binary texts, in which lots of rules were defined to classify scanned document images into different types of blocks [37, 38], such as height, aspect ratio, density, perimeter, perimeter/width ratio, transitions of white-to-black pixels, etc. In [39], the authors focused on detecting handwritten and machine printed texts from the noisy scanned document images. Trained Fisher classifiers are adopted to identify texts from noise based on selected features. These features

(a) Comparison of texts from a natural scene text image (left) and a SCI (right); two corresponding amplified text strings are shown in the middle image.

(b) A text string on document image

(c) A text string on SCI

Figure 2.2: Comparison of some texts on the three kinds of images. In (b) and (c), text strings are amplified to show the detail information.

are geometrical structure of texts derived from Markov Random Field approach. In [43], the morphological grayscale top-hat and bottom-hat filter operators were firstly applied to enhance the document images. Then, horizontal and vertical gradients of each block on the enhanced images are calculated, based on which blocks are classified into text blocks and nontext blocks. The authors of [44] aimed at labeling different components in a document image, such as title, headings, equations, tables and figures. Fuzzy rules were learned from a large set of manually labeled connected components, and then were used to identify different components. In [4], a novel text segmentation algorithm was proposed to improve the segmentation accuracy of the MRC-based document compression. The new algorithms combined two parts for segmentation: cost optimized segmentation (COS) and connected component classification (CCC). The COS is a blockwise segmentation and the CCC refines the initial segmentation using a Markov random filed features. The ideas of these methods cannot be directly applied to SCIs, due to the differences between these two kinds of images. In our experiment, we

implement two state-of-the-art text detection methods [4, 43] designed for document images to test their performance on detecting texts in SCIs in Chapter.3.

Recent years have witnessed the emergence of specialized algorithms for segmenting SCIs. These methods are basically block-based, exploiting properties of text blocks (e.g., color, gradient, transform coefficients etc.) so as to differentiate pictorial and textual blocks. In [8], one SCI is firstly segmented into textual and pictorial blocks according to the color number N of each 16×16 macroblock (if $N > 32$, classified as pictorial block; otherwise, classified as textual block). Then, four kinds of shape primitives(i.e., isolated pixels, horizontal lines, vertical lines and rectangles) are extracted to refine the segmentation. In [10], color and gradient information is combined to decompose a SCI to four categories of 16×16 macroblocks: smooth, text, hybrid and picture blocks. The work in[13] segments SCIs based on the standard deviation of three high-frequency sub-bands of wavelet decomposition. Pan et al. [12] analyzed three visual attributes of textual blocks, i.e., color histogram, gradient distribution and adjacent pixel relationships, based on which, 16×16 macroblocks are decomposed into pictorial and textual blocks. In these block-based methods, the exploited properties are not efficient enough to classify all blocks. Most importantly, these methods do not take into account the varied text scales and orientations. The lack of adaptability not only breaks the integrality of characters, which heavily threatens the identification of texts, but also results in over-connection of textual regions as to be illustrated in Chapter 3.

2.3 Coding Algorithms for Screen Content Images

As introduced in previous sections, due to the special characteristics of texts, existing standard coding methods, such as JPEG, JPEG2000 and H.264 coding, cannot encode the SCIs effectively. Hence, most of existing approaches for SCI compression are segmentation-based, trying to decompose a SCI into different parts or layers, and then to encode these parts separately [10, 12, 45, 46].

A single coding method that simultaneously meets the requirements for both textual and pictorial region compression is elusive. This phenomenon also occurs when using the standard coding methods (i.e., JPEG, JPEG2000 and H.264 intra coding) to encode document images. Compression of document images have been deeply studied by now and lots of typical methods have been proposed to handle this problem. One

Figure 2.3: Illustration of MRC mode for document compression. The foreground and background layers are encoded through standard encoding algorithms (e.g., JPEG and JPEG2000), and the binary mask is encoded by using JBIG or JBIG2.

of the most classical method is the Mixed Raster Content (MRC) method [47, 48, 49], which is initially proposed for compressing compound images containing binary text and continuous-tone images. The most basic MRC mode is shown in Fig. 2.3: it decomposes one image into three layers (foreground, background and binary mask layers). The binary mask indicates the classification of each pixel in the image by 1 or 0. Generally, textual regions are regarded as foreground layer, while pictorial regions are classified to background layer. All layers are then compressed independently using different encoders. The foreground and background layers are compressed using standard encoding algorithms, such as JPEG or JPEG2000, while the binary mask is encoded by using *Joint Bi-level Image Experts Group* (JBIG) or JBIG2 [50]. Moreover, different sub-sampling rates and compression ratios are set to foreground and background layers because of their different features. Typically, the foreground layer is compressed more aggressively than the background layer due to the less color transition in the foreground layer. Lots of MRC based methods were proposed, among which the most representative one is the DjVu [51, 52]. The DjVu achieves good performance in compressing document images that contain texts with binary or limited colors. The benefits of this method will decrease as the increasing of complexity of text color transitions. The reason is that, firstly, as the increasing of text color number, it requires more bits to encode the foreground layer, otherwise the quality of textual regions cannot be ensured.

Furthermore, it increases the difficulty to segment out the foreground layer from the background layer. The classification error can cause distortion in both layers.

Therefore, many methods are then proposed to improve the compression performance for SCIs. In [45], after a simple segmentation, different amount of bits are allocated to textual and pictorial parts through adjusting respective quantization scales in JPEG coding framework (i.e., larger scale for pictorial part and smaller scale for textual part). However, a small quantization scale for textual parts would result in the illegible edges of text, since the high-frequency components in DCT coefficients of textual parts reflect the detailed information of edges in text. Ding *et al.* [10] divided the screen image into four categories of blocks, and designed four different coding algorithms for each category. In this method, the text coding algorithm, which quantizes the colors of textual blocks to four major colors, is ineffective to code the textual components with numerous colors in the screen image. Ding *et al.* [46] introduced color quantization as an alternative mode in the H.264 intra coding framework. The color quantization is a recommendable way to represent limited colors in the textual regions. However, it cannot be used to compress various color scales in the textural part. In [43], the authors proposed a multi-dimensional multi-scale parser (MMP) based encoder for scanned compound images. The encoder is adaptive to textual and pictorial parts. However, the problem with this scheme is the high computational complexity. In [12], one SCI is decomposed into two separate layers. The JPEG and the Portable Network Graphics (PNG) coding algorithm are employed to encode the pictorial and textual layers respectively. Unfortunately, the pixel-domain quantization for textual layer has to be manually controlled according to different bit-rate demands.

Other studies try to combine several compression methods to form separated modes and use the Rate-Distortion Optimization (RDO) criterion to select the best mode for each block in a SCI to achieve the best coding performance. Wang *et al.* [53] proposed a united compression method which integrates the PNG and gzip algorithm into H.264 hybrid coding scheme. In [9], two new intra modes are designed for text compression except the intra modes of H.264. Each block in the screen image chooses its coding mode under the RDO criterion. These combination-based algorithms can achieve much better coding performance compared with the segmentation-based ones. However, the mode selection by RDO is time-consuming and resource-demand.

The importance of SCI compression has been recognized by more researchers from the last two years. At the beginning of 2014, the MPEG/VCEG called for proposals to efficiently compress screen content image/videos with high visual consistency with *Human Visual Systems* (HVS) as an extension of the HEVC standard [18]. Inspired by the competition, many algorithms for screen content coding have been reported since then [19, 20, 21, 22, 23, 24]. In [19], a new coding mode is proposed for HEVC intra coding, in which six edge modes inside a block are defined based on intra prediction directions, corresponding to edge positions. Then, the best edge mode is selected out via a rate distortion optimization. The best edge mode is finally integrated into HEVC framework. The authors in [20] proposed a hybrid encoding approach for HEVC framework. The base-layer is encoded by the conventional HEVC, and the enhanced-layer is handled by using the Golomb based binarization and context-based arithmetic coding. In [22], the authors used non-square prediction unit to replace the square coding unit in intra block copy (IntraBC) prediction, to improve the coding performance for screen content videos. In [23], a non-transform coding scheme is proposed, which can be incorporated into the HEVC framework. In this scheme, screen contents are separated into two parts: color component and structure component. Two coding modes are then designed to encode these components, exploiting the directional correlation and non-translational changes in screen sequences.

2.4 Perceptual Quality Assessment of Screen Content Images

When processing images, various distortions during acquisition, processing, compression, transmission and reconstruction may occur, resulting in degradation of visual quality of these image. Since the processed images are finally presented to human beings, the best manner to measure the visual quality is subjective judgement by human beings. However, subjective testing is generally time and effort consuming, especially for a large amount of images. Hence, the objective metric is much desired to measure the visual quality of images with high consistency with the human visual perception. Generally, the subjective scores are served as ground truth to evaluate performance of the objective metrics.

As to the objective metrics, most existing methods are known as *Full Reference* (FR) manners, in which original reference images can be obtained. In many practical scenarios, however, the reference images are not available. *Reduced Reference* (RR) and *No Reference* (NR) manners are then proposed, using few extracted features of original image or no prior information. According to image contents, objective IQA methods can be generally classified into the following three categories: 1) *Natural Image Quality Assessment* (NIQA); 2) *Document Image Quality Assessment* (DIQA); 3) *SCI Quality Assessment* (SIQA). Since it is impossible to get the original clean document image, almost all DIQA metrics are in NR manner. All the three kinds of manners for NIQA have been widely investigated, while SIQA just attracts researchers' attention in last few years.

NIQA has been studied tremendously during the last decades with lots of algorithms being proposed [30, 54]. Several image quality assessment databases have been constructed [55], such as LIVE [56], IVC [57], TID [58], CSIQ [59], LIVE multiply distorted image database [60], LIVE 3D image database [61], etc. When constructing image quality assessment databases, the visual quality of each test image is obtained via subjective judgement by a lot of subjects. Several subjective testing strategies have been recommended by the *International Telecommunication Union* (ITU) [62], such as the *Single Stimulus* (SS), the *Double Stimulus Impairment Scale* (DSIS) and *Pair-wise Comparison* (PC) methods. Each methodology has its own advantages in certain application scenarios [63, 64, 65], for example, DSIS is recommended to measure the robustness of systems while SS is generally adopted to evaluate large amount of distorted images without reference images. The PC method has been proven to be less annoying for subjects, and has less variation for images with highly apparent distortions [66]. However, one severe disadvantage of PC is that it entails a lot of pair comparison, especially when the number of test images is large, which poses a heavy burden to subjects. Hence, many uncomplete PC methods are proposed [65, 66] to induce less effort and labor to human observers without significantly sacrificing accuracy and reliability of comparison results.

In the subjective testing, general viewing conditions should be strictly in control, including the viewing environment, the configuration of display monitor, the selection of subjects, etc. In the [62], several recommendations are given. When building IQA databases, it is also important to select test images and related distortion types

according to the goal of the assessment problem. ITU also clearly provides the recommendation for the selection. In existing image quality assessment databases, natural images with varied styles are collected and distorted by several distortion types. For example, in the LIVE database, total 1011 images are generated, including 29 reference and 779 distorted images. Five kinds of distortions (i.e., Gaussian blur, white noise, bit errors, JPEG and JPEG2000 compression) are applied to the reference images at different degradation levels. In the CSIQ database, six different types of distortions are applied to 30 original images, including JPEG, JPEG2000, global contrast decrements, additive pink Gaussian noise and Gaussian blurring, which results in 866 distorted versions of reference images. In the TID database which contains 25 reference images and 3000 distorted images, more distortion types are involved (total 24 types), such as spatially correlated noise, quantisation noise, contrast change, JPEG transmission errors, chromatic aberrations, etc. The increased distortions bring much more challenges for designing effective IQA metrics. Besides, many other kinds of images have been increasingly studied, such as 3D images and *High Dynamic Range* (HDR) images, and lots of related databases [61, 67] are then generated for the quality assessment study.

Based upon the above mentioned databases, various classical FR NIQA methods, such as SSIM [68, 69], IFC [70], VIF [71], VSNR [72] FSIM [73], MAD [74], GSIM [75] and GMSD [76], have been proposed to objectively assess the quality of distorted natural images. In [68], three kinds of similarity measures between reference and distorted images are calculated and linearly combined to compute the final quality score of the distorted image. These three measures taken into account three apparent features of images, i.e., illuminance, contrast and structure, which can reflect the image quality change caused by various distortions. A visual information fidelity measure (i.e., VIF) is reported in [71], quantifying the loss of image information in the distortion procedure via proposed distortion model and exploring the relationship between image information and visual quality. In FSIM [73], two kinds of features, i.e., phase congruency and image gradient magnitude, are employed to compute the feature similarity between reference and distorted images. The phase congruency is contrast invariant and the gradient magnitude is adopted as a complementary item in characterizing the local image quality. In [74], the authors advocate to use multiple strategies in HVS to evaluate image quality, i.e., detection-based and appearance-based strategies. To estimate the detection-based distortion in high-quality images, a local luminance and

contrast mask is firstly designed. The appearance-based distortion in low quality images are then measured by computing changes in local statistics of spatial frequency components. These methods have achieved high consistency with human perception as viewing distorted natural images. Besides, many RR NIQA [77, 78, 79, 80] and NR NIQA metrics [81, 82, 83, 84, 85] are also reported.

Document Image Quality Assessment (DIQA) has also attracted more attention in the research community recently due to the increasing requirements of digitization of historical or other typewritten documents [86]. Many document image databases [87, 88, 89, 90] are released, based on which various DIQA methods have been proposed [91, 92, 93]. The effectiveness of the DIQA methods is usually evaluated by the *Optical Character Recognition* (OCR) accuracy calculated by the OCR software. Generally, OCR engines aim to measure the accuracy of algorithms in enhancing document images, rather than the visual quality of processed images. The document images in these databases mainly consist of gray-scale or binary texts, without pictures. Most of these document images suffer from degradations related to the environment, e.g., paper aging, stains, carbon copy effect and reader annotations. Almost all the DIQA methods are designed in no-reference manner, and are implemented at the character (or string) level. For example, in [93], a new document image quality metric is designed by adopting machine learning approaches. For a given distorted image, the proposed method can estimate the degradation level by using a neural network *Multi-Layer Perception*(MLP) regression model. However, this metric is designed for binary document images that have been preprocessed by some binarization methods. Besides, The regression model is trained based on character-level features of document images. The character-level visual quality is much different from image-level visual quality, which limits its versatility in quality assessment of document images.

The topic of *Screen Image Quality Assessment* (SIQA) remains relatively un-explored. Obviously, the DIQA methods cannot be adopted to evaluate the visual quality of SCIs directly, since SCIs include pictorial regions besides textual regions and have no the aforementioned environment-related degradations. The NIQA metrics cannot be directly applied to evaluate the quality of distorted SCIs either, since the statistical features of SCIs are different from those of natural images [8, 31], especially for the textual regions. We provide some natural, text and screen image examples in Fig. 2.4.

2.4 Perceptual Quality Assessment of Screen Content Images

(a) image1

(b) image2

Required activity	Key words
Make a judgment based on established criteria Design or produce something new from component parts Break material down into component parts; perceive interrelationships or hierarchy of ideas Use a concept or principle to solve a problem	Appraise, evaluate, justify, judge, which is better... and why? Design, construct, develop, formulate, imagine, create, change, discuss ideas Differentiate, compare/contrast, distinguish x from y, how does a relate to b?, why does x work? What caused...? ? Apply, solve, show, make use of, modify, demonstrate, calculate, compute
Explain or interpret the meaning Remember facts, terms, concepts, definitions, principles Application Knowledge	Explain, predict, interpret, infer, summarize, paraphrase, convert, translate, give examples, account for Define, list, state, identify, label, name, who? when? where? what?

(c) text image1

(d) text image2

(e) screen image1

(f) screen image2

Figure 2.4: Examples of natural images, textual images and screen content images.

The statistical differences of natural images and SCIs can be measured in terms of naturalness and activity.

The *naturalness* map of an image $I(i, j)$ can be calculated as follows [94]:

$$N'(i, j) = \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + 1} \quad (2.1)$$

where

$$\mu(i, j) = \sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} I(i+k, j+l) \quad (2.2)$$

and

$$\sigma(i, j) = \sqrt{\sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} [I(i+k, j+l) - \mu(i, j)]^2} \quad (2.3)$$

are the local mean and deviation. $i \in \{1, 2, \dots, m\}$ and $j \in \{1, 2, \dots, n\}$ denote spatial indices; m and n represent the image dimension. We compute the distribution of coefficients in $N'(i, j)$. The distributions of the example images are shown in Fig. 2.5. It can be observed that the coefficients of natural images follow a Gaussian distribution.

Figure 2.5: Distribution of values in the naturalness maps of the example images.

In other words, the naturalness of a natural image is high, as demonstrated in [94], while for textual or screen images, the distributions vary greatly. For textual images with textured background (e.g, (c) and (d) in Fig.2.4), the distribution is fluctuated; for screen image consisted of texts and pictures (e.g, (e) and (f) in Fig. 2.4), a sharp pimpling appears.

As to the activity analysis, we utilize the *Block Activity Measure* (BAM) reported in [95]. The activity of a local block $\{b_{ij}\}_{i=1,\dots,m,j=1,\dots,n}$ is calculated as follow.

$$\mathcal{BAM} = \frac{\alpha\sqrt{\mathcal{V}_1} + (1 - \alpha)\sqrt{\mathcal{V}_2}}{m \times n} \quad (2.4)$$

Figure 2.6: Local activity value distributions for natural and textual images.

where

$$\begin{aligned} \mathcal{V}_1 &= \sum_{i=1}^m \sum_{j=1}^{n-1} (b_{i,j} - b_{i-1,j+1})^2 + \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (b_{i,j} - b_{i+1,j+1})^2 \\ \mathcal{V}_2 &= \sum_{i=2}^{m-1} \sum_{j=1}^n (b_{i-1,j} - b_{i+1,j})^2 + \sum_{i=1}^m \sum_{j=2}^{n-1} (b_{i,j-1} - b_{i,j+1})^2 \end{aligned}$$

V_1 is the sum of the 1-distance down-left diagonal and down-right diagonal variations, V_2 is the sum of the 2-distance horizontal and vertical variations, and $\alpha \in [0, 1]$ is weight factor to indicate the percent of variances.

We compute BAM values for 10,000 textual blocks and 10,000 natural blocks with the size of 16×16 from natural and textual images. The distributions of local activity values in the natural images and textual images are shown in Fig.2.6. We can see that the activity values of textual blocks are larger than those from the pictorial blocks, which confirms that a textual image has sharper variation among neighboring pixel values than a natural image.

Hence, the NIQA metrics are not applicable to evaluate the quality of distorted SCIs due to the statistical differences between natural and textual images. In this thesis,

we firstly study the subjective quality of distorted SCIs, and then further investigate the applicability of several state-of-the-art NIQA methods to distorted SCIs. Finally, a specific metric is proposed to objectively evaluate the visual quality of SCIs based on the in-depth analysis of the subjective data for SCIs.

Chapter 3

Scale and Orientation Invariant Text Segmentation

As introduced in Chapter 1, text segmentation is a pivotal step for subsequent processing in the SCIs related applications, such as compression, transmission, retrieval, recognition and quality evaluation of SCIs, and should be further studied to improve segmentation accuracy. To this end, we propose a coarse-to-fine framework for segmenting texts of arbitrary scales and orientations in SCIs. In the coarse stage, the *Local Image Activity Measure* (LIAM) is designed based upon the variation distribution of characters, to highlight the difference between textual and pictorial regions. This stage outputs a coarse textual layer including textual regions as well as a few pictorial regions with high activity. In the fine stage, a *Textual Connected Component* (TCC) based refinement is proposed to eliminate the survived pictorial regions. In particular, a *Scale and Orientation Invariant Grouping* (SOIG) algorithm is proposed to adaptively generate TCCs with uniform statistical features. The Minimum Average Distance (MAD) and morphological operations are employed to assist the formation of candidate TCCs. Then, three string-level features (i.e., shape factor, color similarity and mean activity level) are designed to distinguish the true TCCs from the false positive ones that are formed by connecting the high activity pictorial components. Extensive experiments show that the proposed framework can segment textual regions precisely from SCIs, while preserving the integrity of texts with varied scales and orientations, and avoiding over-connection of textual regions.

3.1 Introduction

It is a challenge to precisely distinguish texts from other components in processing SCIs due to its unique nature: 1) text characters appear in various styles (i.e., sizes, orientations, colors, etc.) without any prior information, and some neighboring characters are connected together; 2) some text characters are embedded or superimposed on complex pictorial regions, increasing the probability of misclassifying text characters; 3) high precision of text segmentation is required for SCIs, because any misclassification of textual regions will introduce artifacts which are easily noticed on this kind of images. Therefore, many researchers have proposed text segmentation algorithms to separate textual components from others in SCIs [10, 12, 13, 33, 34, 35, 45, 46, 96]. In these methods, block-based or character-level objects are commonly assumed to detect texts. However, the block-based or character-level objects cannot extract reliable features for detecting all texts. The assumption may destroy the integrality of texts severely, especially for the large or connected characters, because block-based objects are not robust to variation of character sizes and the character-level features, e.g., stroke width and aspect ratio, cannot achieve a high uniformity for connected characters. String-level textual objects are also introduced in some proposals [3, 4, 36, 44]. As fixed connection scales are assumed to group adjacent characters, these methods result in over-connection¹ and over- or under-segmentation² problems. Furthermore, one commonly used constraint is that, text strings are assumed to be distributed in horizontal or rough horizontal directions.

In this chapter, we propose a novel framework for segmenting texts of arbitrary scales and orientations from SCIs, as shown in Fig. 3.1. Basically, the approach is a coarse-to-fine framework. In the coarse stage, a *Local Image Activity Measure* (LIAM) algorithm is proposed to measure activity values of input images. With the activity map, most of pictorial regions of low activity values can be eliminated, leaving only textual and pictorial regions with high activity values in a coarse textual layer. The

¹Over-connection: text characters are connected with their surrounding high activity pictorial components, forming irregular connected components, and this may result in the misclassification of characters.

²Over-segmentation: characters belong to a same string are separated to isolate components. Under-segmentation: characters in different strings are connected to form a textual component.

Figure 3.1: The proposed coarse-to-fine text segmentation framework. Three main contributions of the proposed method are highlighted by filling.

second stage is to construct *Textual Connected Components* (TCCs) in a scale and orientation invariant manner based on the coarse textual layer. Towards this goal, a *Scale and Orientation Invariant Grouping* (SOIG) algorithm is proposed to estimate the connection scales for characters of arbitrary sizes and orientations, and adaptively connect characters with similar traits into TCCs. The generated TCCs are with uniform statistical features, which can facilitate the text verification that follows. We finally design three criteria for TCC verification, thereby separating texts from pictures exactly. Furthermore, we demonstrate that precise text segmentation can improve the compression performance of SCIs.

To be more specific, our proposed text segmentation framework can segment textual regions from SCIs with the following advantages:

- The proposed framework is robust to detect texts with arbitrary scales and orien-

tations in SCIs with clutter background, which benefits from the proposed SOIG algorithm and the string-level feature extraction and identification.

- The proposed algorithm can preserve the integrality of text characters, avoiding over- or under-segmentation of textual regions, to favor the processing that follows (e.g., to maintain the readability of characters when retrieving texts or ensure the visual quality of textual regions when compressing SCIs).
- For characters embedded or superimposed on complex pictorial regions, over-connection can be well addressed to reduce the misclassification of text characters and pictorial regions.

The remaining of this chapter is organized as follows. The LIAM-based coarse partitioning stage is presented in Sec. 3.2, while the TCC-based refinement is proposed in Sec. 3.3. In Sec. 3.4, we report the experimental results, followed by summary drawn in Sec.3.5.

3.2 Local image activity measure (LIAM) based coarse partition

Given an input of SCI, the coarse stage tries to remove most pictorial regions to favor the following TCC generation and verification. The most prominent difference between textual and pictorial components in SCIs is the distribution of pixel variations. Compared with pictorial components that are relatively smooth, the textual ones usually have more intensive variations, such that they can be roughly separated according to the variation distribution. We hereby propose a new algorithm, LIAM, to measure the local pixel variations in the image, such that the difference between the textual and pictorial components in terms of activity levels is highlighted.

In the proposed LIAM algorithm, four kinds of activity measures for a pixel are considered, as shown in Fig. 3.2: 1-distance variation in horizontal and vertical directions (D1-HV), as well as in diagonal directions (D1-Diag), 2-distance variation in horizontal and vertical directions (D2-HV) and in diagonal directions (D2-Diag). The 1- and 2-distance variation¹, which are generally used to detect the first- and second-order tex-

¹If the chessboard distance between two pixels is k , we term such distance as k -distance, and the variation between them is called k -distance variation.

Figure 3.2: Four activity measures: D1-HV, D1-Diag, D2-HV and D2-Diag.

ture information in images, are adopted to select nearby pixels with large variations, due to the intensive color transition of character edges in SCIs. To show the effect of each measure on the activity computation of texts with arbitrary sizes and orientations, we have collected 100 textual images containing texts of various sizes and styles, and then rotated these images in 12 directions ($-90:15:90$). The overall activity value of an image is computed as the average activity value of all pixels. The distribution of activity values of all the resultant 1300 images in each measure is shown in Fig. 3.3. It can be seen that the D1-Diag achieves the highest values for texts with slight rotation angles, such as in $(-15^\circ, 15^\circ)$, $(-90^\circ, -75^\circ)$ and $(75^\circ, 90^\circ)$, while the D2-Diag gets the highest values for greatly skewed texts with rotation angles in $(-75^\circ, -15^\circ)$ and $(15^\circ, 75^\circ)$.

Figure 3.3: Average activity values of tested images using different measures.

Hence, in our computation of LIAM, we combine D1-Diag and D2-Diag to calculate the activity maps of input images, so as to enhance the activity difference

between texts and pictures of arbitrary orientations. The activity value of a pixel $(p_{i,j}), i = 1, \dots, M, j = 1, \dots, N$ is defined as

$$\mathcal{L}(p_{i,j}) = \alpha \mathcal{V}_1(p_{i,j}) + (1 - \alpha) \mathcal{V}_2(p_{i,j}) \quad (3.1)$$

where

$$\begin{aligned} \mathcal{V}_1(p_{i,j}) = & (p_{i,j} - p_{i-1,j-1})^2 + (p_{i,j} - p_{i+1,j+1})^2 + \\ & (p_{i,j} - p_{i-1,j+1})^2 + (p_{i,j} - p_{i+1,j-1})^2 \end{aligned}$$

and

$$\mathcal{V}_2(p_{i,j}) = (p_{i-1,j-1} - p_{i+1,j+1})^2 + (p_{i-1,j+1} - p_{i+1,j-1})^2$$

are the 1- and 2-distance diagonal variations, respectively. The parameter $\alpha \in [0, 1]$ is served as a weighting factor to tune the combination of D1-Diag and D2-Diag. When α is closing to 0 (resp. 1), the computed activity values of greatly skewed texts are increased (resp. decreased), while activity values of slight skewed texts are decreased (resp. increased). It is demonstrated that $\alpha \in [0.3, 0.5]$ gets the best performance in our experiments. Meanwhile, activity values of 1300 pictures (not including texts) are also computed by LIAM to illustrate the difference. It is observed from Fig. 3.3 that the average activity values for textual images (LIAM-text) are much higher than those for pictures (LIAM-picture).

With the proposed LIAM, a final normalized activity map is obtained. Based on the activity map computed by LIAM, a local adaptive binarization [97] is employed to calculate the binary image $\mathcal{J}_{bin}$ with automatically determined thresholds.

$$\mathcal{J}_{bin}(i, j) = \begin{cases} 0 & \text{if } \mathcal{L}(i, j) \leq T(i, j) \\ 1 & \text{otherwise} \end{cases} \quad (3.2)$$

where $\mathcal{L}(i, j)$ is the activity value at the position (i, j) , and $T(i, j)$ is a threshold at (i, j) and is computed as follows.

$$T(i, j) = m(i, j)[1 + \rho(\frac{\partial(i, j)}{1 - \partial(i, j)} - 1)] \quad (3.3)$$

where $m(i, j)$ is the local mean and $\partial(i, j) = \mathcal{L}(i, j) - m(i, j)$ is the local mean deviation. ρ is a bias which can control the level of adaption varying threshold value. Its range is

3.2 Local image activity measure (LIAM) based coarse partition

(a) A screen content image

(b) LIAM activity map

(c) Binarization of the LIAM map

Figure 3.4: LIAM activity map and its binarization result for one SCI.

[0,1]. As shown in Fig. 3.4 (a) and (b), a SCI and its LIAM activity map are presented. We can see that the LIAM computes high activity values for textual components, and eliminates lots of pictorial components. Actually, as shown in Fig. 3.4 (c), J_{bin} can be considered as an indexing map of the coarse textual layer, in which the corresponding binary values of texts is 1. The coarse textual layer with impurities (i.e. a few pictorial components with high activity) will be further classified in the following refinement procedure.

3.3 Textual connected component (TCC) based refinement

In the refining stage, the target is to remove fake texts from the coarse textual layer, resulting in a cleaner textual layer. The key idea is to construct basic textual objects with uniform features that are scale and orientation invariant. These features are unique to textual objects, such that textual regions can be distinguished from pictorial regions effectively. To this end, we propose a *Scale and Orientation Invariant Grouping* (SOIG) algorithm to construct the desired basic textual objects.

Text characters in SCIs are randomly distributed with various scales and orientations, but characters in one text string usually have similar traits (e.g., color, scale, orientation, etc.). If these characters are connected together, more reliable features can be extracted from the generated string-level *Textual Connected Components* (TCCs), such as color histogram and edge intensity. These features are much stronger than those extracted from single characters to filter fake texts out. The proposed SOIG algorithm can assemble separated characters into TCCs with high adaptivity to their scales and orientations. But in some local regions, characters of small sizes are already connected or very close to each other. It is unnecessary to apply SOIG to these characters, since they can be easily and effectively connected using one fixed small scale. Hence, an image *closing* operation with the minimum scale is firstly applied to the coarse textual layer before SOIG. This operation can not only reduce the complexity of SOIG, but also avoid the under-segmentation among some adjacent textual strings. The flowchart of the TCC-based refinement is given in Fig. 3.5. We first extract and identify TCCs of small sizes from the coarse textual layer. The proposed SOIG algorithm is then applied to adaptively connect the remaining isolated components of large sizes, and

over-connected components are separated through intersection detection and separation. Finally, all the connected components are verified by the defined TCC verification criteria. The final textual layer is obtained by mapping the TCCs back to the original input image.

Figure 3.5: Flowchart of the proposed TCC-based refinement, including the three key steps highlighted by color filling.

3.3.1 Extraction of TCCs with Minimum Scale

In the binary image $\mathcal{I}_{bin}$ resulting from the coarse partition, characters of small sizes have been already partially connected or very close to adjacent characters, while others are still distributed in certain distance. In this step, we focus on the former cases, and leave the latter ones for further processing. The morphological closing operation [98] with a minimum scale structural element (i.e. a kernel se) is employed to construct

candidate TCCs:

$$\mathcal{J}_{close} = (\mathcal{J}_{bin} \oplus se) \ominus se \quad (3.4)$$

where $\mathcal{J}_{close}$ is an logical image which represents all formed components in a binary fashion, $\oplus$ and $\ominus$ represent dilation and erosion operation respectively, and se is defined as 3×3 matrix with $se_{i,j} = 1$, $(i, j = 1, 2, 3)$, which is generally regarded as the minimum scale for closing. Through the closing operation, small characters that are partially connected or very close to adjacent ones are completely connected. To screen out TCCs from all the connected components, the proposed TCC verification criteria¹ are utilized to verify all connected components. The components satisfying the criterion (i.e. TCCs) are moved onto a final closed image $\mathcal{I}'_{close}$, and others are still preserved in the closed image $\mathcal{J}_{close}$.

3.3.2 Scale and Orientation Invariant Grouping (SOIG)

For the remaining components in $\mathcal{J}_{close}$, using fixed connecting scales cannot construct TCCs with uniform features, as text characters are usually with multiple sizes. And, distances between each two characters or distances between each two text strings are usually changed along with character size variation. Moreover, the varied orientations also bring difficulties in determining the connecting scales. In our framework, the SOIG algorithm is proposed to adaptively connect separated characters based on their sizes and orientations.

As shown in Fig. 3.6, where the characters' centroids are marked by red circles, it can be observed that each character is detected as an isolate component. Generally, no matter how a text string is skewed, the intra-string distance d^c between characters is smaller than the inter-string distance d^w . In a same string, the intra-distance between each two characters are almost the same. Furthermore, the distance d^c and d^w generally ascend with increasing character sizes. Additionally, the distance between two strings in different alignments, d^l , is larger than d^c . Hence, giving the mean intra-string distance between characters in each string as a connecting scale, which is actually a metric invariant to scale and orientation, we can adaptively connect these characters properly and guarantee the consistency of geometrical features of constructed TCCs.

¹To maintain the fluency of this chapter, we will describe the verification rules in Sec. 3.3.3.

Figure 3.6: Illustration of intra-string and inter-string distance between text characters.

We hereby propose *Minimum Average Distance* (MAD) to calculate the mean intra-string distance between each two adjacent components. Given two point sets (characters in our case), A and B , we first define the *Directional Average Distance* (DAD) from A to B as follows:

$$\mathcal{H}_{K_1}(A, B) = \frac{1}{K_1} \sum_{b=1, b \in B}^{K_1} \min_{a \in A} \|a - b\|_2 \quad (3.5)$$

where a and b indicate the points in sets A and B separately, and $K_1 = \lceil \gamma \times |B| \rceil$ ($0 < \gamma \leq 1$) in which $|\cdot|$ denotes the size of a point set, $\lceil \cdot \rceil$ is the rounding down operation, and γ is sampling frequency. We uniformly select $\lceil \gamma \times |B| \rceil$ samples from B to simplify the computation. Similarly, the DAD from B to A is defined as $\mathcal{H}_{K_2}(B, A)$, and $K_2 = \lceil \gamma \times |A| \rceil$ ($0 < \gamma \leq 1$). We then give the definition of MAD as follows.

$$\mathcal{D}(A, B) = \min\{\mathcal{H}_{K_1}(A, B), \mathcal{H}_{K_2}(B, A)\} \quad (3.6)$$

Essentially, the MAD represents the mean distance between the closest edges of two characters. As an example, Fig. 3.7 illustrates the calculation of the DAD from A to B .

The flowchart of the SOIG algorithm is shown in Fig. 3.8, and more details are described in **Algorithm 1**. In the traditional rectangle bounding box, standard rectangles are widely adopted to measure the sizes of isolate components. The rectangle bounding box is only suitable to estimate the size of horizontally or vertically distributed text strings, and more false area will be involved for text strings with other directions. Hence, we employ the minimum area rectangle [99] to fit the components with better adaptability to orientation. The minimum area rectangle can adaptively adjust the rectangle's direction according to the text string. The minimum area of the rectangle is regarded as the area of a component ($area(\cdot)$), and the minor length of the

Figure 3.7: Calculation of the DAD between two point sets: A and B . $d(a_i, b_j)$ computes the nearest Euclidean distance from A to b_j .

rectangle is regarded as the width of the component ($l_{min}(\cdot)$). Given a set of isolate components C and their centroid set O , we denote the component with minimum index (e.g. t) by C_t . First of all, three rules are defined to find the nearest neighbor C_p to C_t ($p \neq t$):

1. the distance between O_t and O_p fits: $d(O_t, O_p) < \beta \times \max(l_{min}(C_t), l_{min}(C_p))$;
2. the area ratio between C_t and C_p should meet: $area(C_t)/area(C_p) < \delta$ or $area(C_p)/area(C_t) < \delta$;
3. when finding one more component C_q to be connected to $\{C_t, \dots, C_p\}$, the angle between the lines $O_t O_p$ and $O_t O_q$ should be less than θ' .

Here, β , δ and θ' are parameters to control the searching. Rule 1) is designed to find the character that is not far away from C_t , so as to avoid involving too much backgrounds. Rules 2) and 3) are set to ensure that characters to be connected are with similar scales and orientations. Hence, the feature uniformity of formed TCCs can be ensured to favor the TCC verification.

If MAD between C_t and C_p is larger than a specific distance, C_t is isolate and needs to be self-connected (Lines 6-9 in Algorithm 1). Otherwise, we use $\mathcal{D}(C_p, C_q)$ to check whether C_q (the nearest component to C_p) is bounded to be combined with C_t

Algorithm 1: Scale and Orientation Invariant Grouping (SOIG) of isolate components

Input: A binary image with a set of isolate components: $C = \{C_1, C_2, \dots, C_n\}$, and corresponding centroids: $O = \{O_1, O_2, \dots, O_n\}$; n is the total number of isolated components.

Output: A set of candidate TCCs: $C' = \{C'_1, C'_2, \dots, C'_m\}$. m is the number of connected components.

```

1 Initialization:  $C' = \emptyset$ ;
2 while set  $O$  is not empty do
3   for  $C_t$  ( $t$  is the minimum index in set  $O$ ), find the component  $C_p$  nearest to
    $C_t$  according to the first two rules;
4   Calculate the MAD between  $C_t$  and  $C_p$ :  $\mathcal{D}(C_t, C_p)$ ;
5   Get the width of  $C_t$  and  $C_p$ :  $l_{min}(C_t)$  and  $l_{min}(C_p)$ ;
6   if  $\mathcal{D}(C_t, C_p) > \beta \times \max\{l_{min}(C_t), l_{min}(C_p)\}$  then
7     remove  $O_t$  from the set  $O$ ;
8     connect  $C_t$  in the distance of  $l_{min}(C_t)$ ;
9     arrange connected  $C_t$  in  $C'$ ;
10  else
11    initialize the mean distance between the isolate components:
     $L_{mean} = \mathcal{D}(C_t, C_p)$ ;
12    find the point  $C_q$  nearest to  $C_p$  according to the defined three rules;
13    if  $C_q$  exists then
14      add the value of  $q$  to the index set  $O' \{t, \dots, p, q\}$ ;
15       $p = q$  ;
16      update the value of  $L_{mean}$ :  $L_{mean} = \frac{(num(O')-1)*L_{mean} + \mathcal{D}(C_p, C_q)}{num(O')}$ ;
17      return to step 12;
18    else
19      calculate the main orientation  $\theta$  of the components of the indices in
      the set  $O'$ ;
20      Connect these components using designed structural kernel  $\acute{K}$  based
      on the  $\theta$  and  $L_{mean}$ ;
21      arrange the connected component into  $C'$ ;
22      remove the indices in  $O'$  from the set  $O$ ;
23    end
24  end
25 end

```

Figure 3.8: The flowchart of the SOIG algorithm. The neighbour searching is implemented based upon the defined three rules. The adaptive computation of θ and L_{mean} is given in detail in Algorithm 1.

and C_p . This iteration is terminated when no more components are supposed to be connected (Lines 11-17). By far, we get an ordered subset of components $\{..., C_p, C_q\}$ that should be connected together with C_t . We finally connect these components according to an orientation-related structure kernel $\hat{K}$ for the components (Lines 19-20). The orientation-related structure kernel is designed for connecting these components as follows, to avoid too much dilation of surrounding irrelevant regions. According to the centroids in O' , we can estimate the main orientation of the connection and quantize it into one of the intervals $0 : 30 : 180$. Then, based on the orientation θ and L_{mean} , the kernel is generated. For example, in Fig. 3.9, if $L_{mean} = 7$ and $\theta = 30^\circ$, the kernel is designed where values inside the central $\lfloor \frac{L_{mean}}{2} \rfloor$ lines in the direction are set to 1; otherwise to 0.

0	0	0	0	0	0	1
0	0	0	0	1	1	1
0	0	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	0	0
1	1	1	0	0	0	0
1	0	0	0	0	0	0

Figure 3.9: An orientation-related structure kernel. Four central lines are marked by dotted lines and the direction of these lines is 30° .

When generating candidate TCCs, for some texts which are closely surrounded by pictorial components, over-connection may appear, which breaks the uniformity of extracted string-level features of TCCs. In these over-connected components, pictorial parts generally have different orientations from textual parts. Accordingly, we employ the morphological operation *skeletonization* to find their skeleton and intersections [36, 100]. Then we can overlay the skeleton over the components and trace each separated components after segmenting the intersections. Finally, each separated component can be regarded as a candidate TCC. Consequently, through our algorithm, text characters are gradually connected and identified, and the TCCs are also put to the final closed image I'_{close} . At last, the final textual layer can be visualized by mapping the closed image I'_{close} to the original input image. Fig. 3.10 shows some intermediate results of an image sample for better understanding the TCC-based refinement. In Fig. 3.10 (a), after the image *closing* operation, adjacent characters of small sizes are connected to form new components on the coarse textual layer, while characters with large sizes are still isolated. All the newly formed components are verified by the defined TCC verification criteria, and the ones satisfying the criteria are moved onto a final closed image. Then, the SOIG algorithm is applied to the remaining components, and the connection result is shown in Fig. 3.10 (b), where the isolated large characters as well as pictorial ones are connected. These newly formed components satisfying the TCC verification criteria are also put on the final closed image. The final textual layer is obtained by mapping the final closed image to the original image, and the segmentation result is marked by green boxes as displayed on Fig. 3.10 (c). It can be observed that all the text strings are segmented without over- or under-segmentation. Moreover, the

integrality of text characters is well preserved especially for the skewed text strings.

(a) Closed image with the minimum scale

(b) Adaptive connection result by SOIG

(c) Final segmentation result

Figure 3.10: Results of intermediate steps in the TCC-based refinement.

3.3.3 Criteria for TCC Verification

Since the textual components are “polluted” by a few pictorial components, these fake texts may also be connected. After aforementioned processes, we obtain a set of candidate TCCs, $\mathcal{T} = \{T_1, T_2, \dots, T_{n'}\}$, where n' is the number of candidate TCCs. These components may have various scales, orientations and aspect ratios (ratio of their major length and minor length). Hence, the desired features should be uniform and unique to TCCs, such that false positive TCCs can be eliminated. Three types of string-level features, i.e. shape factor, color similarity and mean activity value, are designed to exactly distinguish true TCCs from false positive ones.

Figure 3.11: Feature analysis of different candidate connected components.

Characters in a text string generally have similar styles, sizes and orientations. Therefore, true TCCs have similar shape after the proposed grouping, while the false positive ones have irregular shapes. Here, the minimum area rectangle is utilized to approximate the area of connected components, rather than the widely used bounding box, as it has better adaptability to varied orientations of components. For true TCCs, most of the minimum rectangle area will be occupied by non-zero pixels, while the area for false positive TCCs is filled by lots of zero pixels, as shown in Fig. 3.11 (a) and (d) respectively. The shape factor of a TCC, $\mathcal{S}(T_t)$, ($1 \leq t \leq n'$), is defined as:

$$\mathcal{S}(T_t) = \frac{area(T_t)}{r_1(T_t) \times r_2(T_t)} \quad (3.7)$$

where $area(T_t)$ represents the sum of non-zero values in T_t , while $r_1(T_t)$ and $r_2(T_t)$ are the major length and minor length of the fitting rectangle of T_t respectively. If $\mathcal{S}(T_t)$ is larger than a threshold ξ_s , T_t is classified as a true TCC.

In order to improve the verification precision, other two characteristics (i.e. color similarity and mean activity level) of candidate TCCs are also utilized. For textual components with smooth background, characters in the same string usually have similar

color distribution and there are only a few major colors. Therefore, for each $T_t \in \mathcal{T}$, we get its corresponding area A_t in the original input image. For example, in Fig. 3.11, for the two components shown in (a) and (d), their corresponding areas in the original image are given in (b) and (e). It is obviously observed that the color histogram of the true TCC (c) is different from the one of the pictorial components (f): the textual component contains a few major colors, while the number of colors in the pictorial component varies drastically.

Hence, for textual components on smooth background, the ratio of pixels occupied by the major colors (see Eq. 3.8 below) is larger than a threshold ξ_c . The major colors are decided as the colors with the first k maximum occurrence times. We calculate the occurrence time of each color in the area A_t , ($1 \leq t \leq n'$) and arrange them in a descending order: $\mathcal{P} = \{P_1, P_2, \dots, P_J\}$, where J is the total number of colors. The ratio $\mathcal{R}_c$ of major colors can be defined as:

$$\mathcal{R}_c(A_t) = \sum_{i=1}^k P_i / \sum_{j=1}^J P_j \quad (3.8)$$

where k is set to 6 if k is larger than J ; otherwise $k = J$.

For textual components superimposed on complex pictorial background, of which the color histograms are background dependent, the major color density may not be a stable feature for TCC identification. We thus consider the mean activity value as another feature to distinguish these textual components. The activity map $\mathcal{I}_{act}$ of the input image has been obtained through the LIAM, as shown in Fig. 3.4 (b). It can be observed that textual regions have more compact distribution than pictorial regions. Additionally, the activity values of textual pixels are generally much larger than those in pictorial regions. According to the connected components $\mathcal{T}$, corresponding regions $E = \{E_1, E_2, \dots, E_{n'}\}$ on the activity map are acquired. The mean activity value shown in Eq. 3.9 of a TCC is generally larger than a threshold ξ_a .

$$\mathcal{M}(E_t) = \frac{1}{k'} \sum_{i=1}^{k'} E_t(i) \quad (3.9)$$

where k' represents the number of non-zero values in E_t .

Based on the designed constraints, a candidate component T_t can be marked as a TCC if:

$$\mathcal{U}_1 \cap (\mathcal{U}_2 \cup \mathcal{U}_3) = 1 \quad (3.10)$$

where

$$\begin{cases} \mathcal{U}_1 : \mathcal{S}(T_t) \geq \xi_s \\ \mathcal{U}_2 : \mathcal{R}_c(A_t) \geq \xi_c \\ \mathcal{U}_3 : \mathcal{M}(E_t) \geq \xi_a \end{cases}$$

With this formulation, a candidate component can be judged as a TCC if it satisfies the rule in (Eq. 3.10).

3.4 Experimental Results

In this section, we firstly compare the proposed LIAM with some other methods to verify its effectiveness in distinguishing texts from pictures. Then, the proposed segmentation algorithm is evaluated on two databases, i.e. the public ICDAR database [31] and a newly designed SCIs database (BCID) [101]. An application of the proposed algorithm on compression of SCIs is demonstrated, and the performance on the BCID database is compared with some state-of-the-art methods.

3.4.1 Effectiveness Analysis of the proposed LIAM

In order to illustrate performance of the LIAM, some artificial SCIs are manually created (e.g., Fig. 3.4 (a)), and the ground truths of textual regions (as shown in Fig. 3.12 (a)) are obtained as binary images. We compare performance of the LIAM with other four methods, i.e. *Spatial Frequency Measure* (SFM)[102], *Image Activity Measure* (IAMx)[103], Canny detector and *Laplacian of Gaussian* (LoG) filter, in extracting candidate textual regions from SCIs. The activity maps derived from SFM, IAMx and LIAM, or edge maps from Canny detector and LoG filter, are all binarized using the same binarization algorithm [97]. The binarized images are shown in Fig. 3.12 for visual comparison (refer [101] for more detailed results). It can be observed from Fig. 3.12 (c) that SFM cannot obtain high activity values for all textual pixels; some edges of texts (especially the ones superposed on pictures) are almost not detected, which leads to the breakage of text integrality. The IAMx can detect textual regions well, as displayed in Fig. 3.12 (d), but it involves more pictorial regions which will increase the complexity to filter texts out in sequential procedures. In Fig. 3.12 (e) and (f), we can see that Canny detector cannot detect the inner structure of small connected characters, resulting in the lost of detailed activity information, while the LoG involves more pictorial

pixels with clear character shape, and as pictorial pixels reducing, character shape will be destroyed gradually. The LIAM can preserve characters with low contrast on the coarse textual layer and eliminate a large part of pictorial components.

Additionally, we also objectively compare the textual segmentation performance of these methods. We compare the binary images with the ground truths of texts at the pixel level in terms of precision (p), recall (r) and F-measure (f). A standard F-measure f is defined as the harmonic mean of the p and r :

$$f = 2 \times p \times r / (p + r) \quad (3.11)$$

where p and r is defined as following:

$$p = TP / (TP + FP), \quad r = TP / (TP + FN) \quad (3.12)$$

TP is the number of correctly detected textual pixels. FP is the number of pixels which are pictorial pixels, but detected as textual ones. FN is the number of pixels which are textual pixels, but misclassified as pictorial ones.

The objective comparison results are shown in Table 3.1. As shown in Table 3.1, the F-measures of Canny detector and LoG filter are quite low, because they are proposed for detecting image edges rather than detecting just textual pixels. The recall of the SFM is low ($r = 0.6136$ while $p = 0.8144$), which indicates that lots of textual pixels are eliminated; the low precision value ($p = 0.7574$ while $r = 0.8243$) of IAMx means that many pictorial pixels are retained on the activity map. The proposed LIAM achieves the highest precision and F-measure values comparing with the four methods, which indicates that the proposed LIAM can erase most of pictorial regions and maintain textual regions well on the coarse textual layer. Note that, the LIAM is proposed for roughly detecting textual regions, and it just detect edge regions of large characters. Hence, the precision and recall values are a little bit low based on the obtained ground truth map. Therefore, further refinement is required to exactly segment textual regions out.

The reason for the less effectiveness of SFM and IAMx comes from their computation of pixel variations. As illustrated in Fig.3.13 (a), the SFM computes variation of a pixel only based on its two closest pixels in horizontal and vertical directions. Hence, the computed values of text edges with low contrast (e.g., texts superposed on pictures) are much smaller than the ones of high contrast texts, and are even lower than those

(a) Ground truth of textual regions

(b) Binarization of LIAM activity map

(c) Binarization of SFM activity map

(d) Binarization of IAMx activity map

(e) Canny detection

(f) LoG filtering

Figure 3.12: Comparison of binarized results of activity maps generated by different methods.

Table 3.1: Comparison of text detection results on artificial SCIs.

Method	Precision (p)	Recall (r)	F-measure (f)
LIAM	0.8376	0.7840	0.8098
SFM	0.8144	0.6136	0.6984
IAMx	0.7574	0.8243	0.7893
Canny detector	0.6846	0.3654	0.4690
LoG filtering	0.6526	0.5115	0.5723

Figure 3.13: Computation of pixel variation in (a) SFM and (b) IAMx.

of pictorial pixels. As a consequence, structure of low contrast texts will be severely destroyed after binarization. The IAMx also measures variations in horizontal and vertical directions, as shown in Fig.3.13 (b). In this way, the variations of textual and pictorial pixels are not highlighted, since text variations are simultaneously strong in various directions (e.g., horizontal, vertical and diagonal directions), while pictorial variations are generally smooth or just strong in certain direction. Therefore, based on the IAMx, more pictorial pixels will be involved, increasing the difficulty to segment texts out.

3.4.2 Text Segmentation On ICDAR Database

The ICDAR database is commonly used for testing text detection algorithms on natural images. In this database, most of characters are horizontally distributed and superimposed on pictorial backgrounds. We test the proposed TCC-based segmentation framework on the ICDAR database, and compare it with some existing proposals [104] which have been involved in the competition organized during ICDAR 2011 for detecting texts from web images [31].

Results for the comparison are shown in Table 3.2 and Fig. 3.14. It is seen from

Table 3.2 that our method (named as TCC-textDetector) achieves the best recall and Harmonic Mean values. The higher recall value indicates that less texts are misclassified as pictorial components. The precision of our method is a little lower because some character-like symbols are classified as characters (e.g., the symbol $\sqrt{\quad}$ in Fig. 3.14 (b)), or some artistic characters superimposed on complex pictures (such as the symbol "\$" shown in Fig.3.14 (b)) are missed. But our method obtains the best overall performance in term of Harmonic Mean. As shown in Fig. 3.14 (a), our algorithm can detect texts of various scales and orientation, avoiding over- or under-segmentation.

Table 3.2: Experiment results on ICDAR database.

Method	Recall	Precision	Harmonic Mean
Textorter	0.6962	0.8583	0.7688
TH-TextLoc	0.7308	0.8051	0.7662
TDM IACAS	0.6916	0.8464	0.7612
OTCYMIST	0.7591	0.6405	0.6948
AlvaroGonzalez	0.7008	0.8923	0.7851
TDM-IACAS	0.6970	0.8535	0.7978
Anthimopoulos	0.8188	0.8735	0.8435
TCC-textDetector	0.8864	0.8146	0.8490

We also evaluate one advanced text detection method [3], which works well for detecting texts in natural scene images in the tested database. The comparison results are shown in Fig. 3.15. It can be observed that the method in [3] can detect large and separated characters on the tested images, but leave out lots of small or connected characters. Our algorithm achieves much better results, which can not only detect all texts correctly, but also avoid under- or over-segmentation of text strings.

The advantage of the proposed algorithm has not been well demonstrated with the ICDAR database, since the ground truths given in the database are bounding boxes enclosed texts which are not appropriate to rotated strings. As shown in Fig.3.14 (c), we can more exactly locate the regions of skewed texts, but the precision and recall are both zero according to the marked ground truth and the evaluation method. Moreover, most of texts in this database is distributed in horizontal direction. We will report more

(a) Text segmentation results of our algorithm for some examples.

(b) Example1: recall: 83.33%, precision: 83.33%, Harmonic mean: 83.33%

(c) Example2: recall: 0%, precision: 0%, Harmonic mean: 0%

Figure 3.14: Experiment results for some samples in ICDAR 2011. In (b) and (c), the left is the ground truth and the right is our result.

Figure 3.15: Comparison of text detection results. The left images are results using the method in [3], and the right ones are results of our method.

experiments on images which are more comprehensive in the following section.

3.4.3 Text Segmentation on the BCID database

Different from the ICDAR database, images in the BCID database include texts of arbitrary scales and orientations. And, texts are mixed with pictures in more comprehensive manners: texts are gathered on smooth background, enclosed by pictures or superimposed on pictures. The text features change greatly, resulting in difficulty to simultaneously detect all texts in SCIs. In this experiment, images in the BCID database are randomly captured from computer screens. The database includes five categories of SCIs, i.e. PDF, webpage, power point (PPT), digital magazine and combined images. Such images are widely used in many applications [1, 2, 6]. 100 testing images are used to evaluate our segmentation algorithm. In these images, the scales, orientations and positions of text characters vary significantly. The BCID database is publicly available now on the website [101].

In our experiments, the proposed TCC-based text segmentation algorithm is compared with other two existing methods for SCIs: Visual Attributes (VA) based method [12] and DWT-based method [13]. Meanwhile, another two methods designed for document images are also implemented, including the GA-based method [43] and the

Haneda's method [4]. Since existing methods are commonly block-based, we evaluate our framework as well as other methods using a block-level evaluation methodology that is defined in terms of precision, recall and F-measure. We first transform our segmentation results to the block level form. In particular, we decompose the final index map (obtained in Sec. 3.3.3 above) into blocks. A $h' \times l'$ window is used to scan the index map, and a block b is labeled as 1 if it satisfies the following criterion:

$$J_{\mathcal{P}}(b) = \begin{cases} 1 & \text{if } \sum_{i=1}^{h'} \sum_{j=1}^{l'} b(i, j) > \min(h', l') \\ 0 & \text{otherwise} \end{cases} \quad (3.13)$$

where $b(i, j)$ represents the value of a pixel at the location (i, j) in the block b . Then we can obtain a segmentation index $J_{\mathcal{P}}$ at the block level.

We then use the following approach to construct the ground truth of the segmentation. We use a window of size $h' \times l'$ (both h' and l' are set to 16 as commonly used in the compared block-based methods) to scan all the images, then manually label all the blocks. The blocks including text characters are labeled as 1, and 0 otherwise; the resultant labeled index $J_{\mathcal{G}}$ is the ground truth of the segmentation. The performance measures p and r can be calculated by using (Eq. 3.11) and (Eq. 3.12) based on the $J_{\mathcal{G}}$ and the $J_{\mathcal{P}}$ at the block level as in [13]. Here, TP is the number of correctly classified textual blocks. FP is the number of blocks which are pictorial blocks, but classified as textual blocks. FN is the number of blocks which are textual blocks, but misclassified as pictorial blocks.

Two example SCIs and their segmentation results by the four methods are provided in Fig. 3.16 and Fig. 3.17. It can be found from these two figures that the proposed method not only locates the textual regions exactly, but also achieves high classification precision for both kinds of blocks. The VA-based method can remove most of pictorial blocks, but it misjudges many texts superimposed on pictorial regions. The DWT-based cannot distinguish the pictorial blocks of high activity from textual blocks; for example, the image border blocks including strong edges or pictorial blocks with complex texture. The GA-based method can detect most of the textual regions. But it also misclassifies some pictorial regions with high activity or surrounding texts. Additionally, the text detection result of Haneda's method on a SCIs is shown in Fig. 3.18. It can be observed that the method can detect large characters superimposed on pictorial regions, but destroy the integrality of small characters severely.

(a) SCI example I

(b) VA-based method I

(c) DWT-based method I

(d) GA-based method I

(e) TCC-based method I

Figure 3.16: Segmentation results of example I by the four methods. The segmented textual layers are displayed in (b)-(e), respectively (the padded pixels are set to 0 in the textual layers).

(a) SCI example I

(b) VA-based method II

(c) DWT-based method II

(d) GA-based method II

(e) TCC-based method II

Figure 3.17: Segmentation results of example II by the four methods. The segmented textual layers are displayed in (b)-(e), respectively (the padded pixels are set to 0 in the textual layers).

Figure 3.18: Text detection of [4] of a tested image (the padded pixels are set to 0 in the segmented layers).

To further verify the overall effectivity of the proposed segmentation algorithm, we also perform quantitative comparison among the methods in terms of precision p , recall r and F-measure f as tabulated in Table 3.3. From Table 3.3, it is shown that our framework achieves the highest F-measure, which means a more precise segmentation for the test images. The reason for the high precision and recall is that the proposed algorithm can not only preserve the integrality of texts with varied sizes and orientations, but also avoid over- and under-segmentation of textual regions. The VA-based method may misjudge both types of blocks for all images. For PDF images, the latter three methods (DWT-, GA- and the TCC-based methods) all perform very well, due to the their simple background. But for other compound images, the DWT- and the GA-based method obtain high recall ratios but relative low precision ratios; in other words, these two methods can detect textual areas well, but classify many pictorial blocks as textual blocks.

3.4.4 Impact of Segmentation Accuracy on SCI Compression

As aforementioned in Chapter 1, SCI segmentation is a critical step in SCI compression framework, and the segmentation accuracy will affect the compression performance. In this section, we will demonstrate that low segmentation accuracy will degrade the coding performance. We perform the experiments on the BCID-database. These SCIs are

Table 3.3: Experimental results on BCID database

	VA-based method [11]			DWT-based method [12]		
images	p (%)	r (%)	f (%)	p (%)	r (%)	f (%)
PDF	89.92	70.74	79.19	93.27	99.01	96.05
PPT	65.34	82.87	73.07	71.32	97.08	82.23
wallpage	79.02	69.49	73.95	75.43	97.92	85.22
webpage	84.76	74.37	79.23	80.34	98	88.29
combined	77.37	81.43	79.35	77.12	99.08	86.73
Average	79.28	75.78	76.96	79.5	98.23	87.71
	GA-based method [10]			Proposed TCC-based method		
	p (%)	r (%)	f (%)	p (%)	r (%)	f (%)
PDF	99.18	99.5	99.34	99.8	100	99.9
PPT	83.65	93.5	88.3	91.16	95.57	92.74
wallpage	80.02	98.73	88.4	94.62	98.17	96.36
webpage	88.28	97.24	92.54	97.63	99.87	98.76
combined	85.4	99.78	92.03	93.89	98.09	95.94
Average	87.31	97.75	92.12	95.42	98.34	96.86

segmented into two layers: a pictorial layer and a textual layer, by using the proposed coarse-to-fine framework as well as other segmentation algorithms (e.g., GA-based, DWT-based and VA-based methods). Then, we encode these images by using one same layer-based compression strategy [11]. In the layer-based compression strategy, a specialised text codec based on basic colours and index map is applied to encode the textual layer, while the pictorial layer is compressed using the H.264 intra coding algorithm.

In our comparison, we put the textual and pictorial blocks resulting from the block-based segmentation (e.g., DWT-based and VA-based methods) on a textual layer and a pictorial layer respectively without indexing. In the layer-based coding algorithm, one crucial step is to fill the segmented regions on both layers to smooth transitions. Some data-filling methods have been proposed [105], where an index map at the pixel level is required to indicate that the value of each pixel is selected from textual or pictorial layer. The filling procedure is time-consuming, and the index map as well as the two layers need to be encoded. Hence, we employ a simple data-filling strategy without an index map. We only fill the padded pixels with value of 255 (white). The original image can be reconstructed by directly subtracting 255 from the overlay of the two layers. By

this way, what we need to do is to code the two layers.

Generally, the human vision is more sensitive to the artifacts in textual regions. Hence, larger quantization steps are applied to the pictorial layer, while quantization steps for the textual layer are smaller when encoding SCIs. Using this compression strategy, text segmentation precision is crucial: for the misjudged textual components, large quantization steps for pictorial layer will obscure the text characters; for pictorial components misclassified as textual regions, much more bits are need to code the basic colors and the index map. Besides, we compare the compression performances with four standard coding algorithms, i.e. JPEG, JPEG2000, H.264 intra coding and DjVu. Different from compressing document images, we don't sub-sample in any layer due to low resolution of the SCIs when performing DjVu.

Compared with the standard coding methods, the segmentation based compression achieves much better visual quality at the same bitrates, which can be observed from Fig. 3.19. From Fig. 3.19 (a)-(f), we can see that severe blocking and blurring artifacts appear on the reconstructed images via JPEG and JPEG2000 compression. On the reconstructed image via H.264 intra coding, the background regions will be affected by the adjacent text characters. The DjVu algorithm reserves texts with clear shape, but deforms pixel transitions of texts as well as pictorial regions, which severely destroy the fidelity of the original image and naturalization of edges. With the help of the proposed segmentation scheme, compression can be performed better: the text characters are clearly displayed while their transitions are well persevered.

The overall comparison of these methods in terms of R-D curves for tested images is shown in Fig. 3.20. It can be observed that the proposed TCC-based method achieves the best PSNR value at the same compression ratio. The DWT-based and the GA-based methods obtain better performance than the VA-based method. It has verified that precise segmentation can improve the coding performance of SCIs. Besides, comparing with JPEG, JPEG2000, DjVu, and H.264 intra, the proposed segmentation based compression is more efficient in coding SCIs. The average PSNR of the proposed segmentation based coding is 5 dB higher than H.264 intra coding, which verified the effectiveness of the segmentation based compression scheme.

Figure 3.19: Visual quality of reconstructed image based on different coding methods. The bitrates for all these images are 0.4 bpp. (f) is based on the proposed segmentation algorithm.

3.4.5 Further Analysis of the Proposed Algorithms

In the proposed coarse-to-fine framework, two algorithms, LIAM and SOIG, are proposed to segment texts in SCIs. As the variations of texts on the selected directions are always larger than that of background regions, LIAM is designed to sensitively capture such variations, guaranteeing the textual regions to be fully detected. The following binarization plays a pivotal role in computing the coarse textual layer. As a global binarization may eliminate the texts with low contrast, we hereby adopt a local adaptive binarization algorithm to maintain the textual pixels, thereby further enhancing the robustness of the coarse partition.

In the SOIG algorithm, one key procedure is the calculation of MAD between two isolated components. To be specific, it calculates the mean distance between the closest edges of two components, which is better than the centroid- and minimum-distance be-

Figure 3.20: Compression results for SCIs.

tween two characters. One connection example is illustrated in Fig. 3.21. The centroid distances are affected by the shapes of characters, which is generally longer than the needed connecting distance. This may increase the under-segmentation on the final textual layer. The minimum distance may result in uncomplete connection between characters that may affect the feature uniformity and result in the over-segmentation of texts. Additionally, unlike the generally used connecting methods in which 4- (or 8-) connectivity or fixed multi-scales are applied to generate basic textual objects at one time, the proposed SOIG algorithm adaptively connects text strings one by one, regardless of their scales and orientations. In this way, over-connection can be eliminated greatly. Moreover, the morphological operation, i.e. skeletonization, is utilized to separate some over-connected components where texts are closely surrounded by pictorial components.

There are three predefined parameters β , δ and θ' in SOIG algorithm to ensure that characters to be connected are with similar scales and orientations, no matter how largely their scales vary and how skewed the orientations become. The settings of the three parameters may affect the feature uniformity of formed TCCs in terms of shape factor, color similarity and mean activity value. For example, if δ becomes larger, characters with large scale difference will be connected together, and then the

Figure 3.21: Connection results using different distance measure.

shape factor of TCCs will be reduced. Given a larger β , more background regions will be involved, which can result in the decrease of the color similarity and mean activity value of formed TCCs. More false positive TCCs will be induced in the above cases. In our experiments, the three parameters are set to 3, 3 and 25° , according to the observations of text characteristics in [106]. Then, we determine the three thresholds ξ_s , ξ_c and ξ_a based upon our extensive experiments on training images in the two databases. In particular, for each training image, all true TCCs and false positive TCCs can be identified by the TCC verification. We then calculate the three features (i.e. shape factor, color similarity and mean activity value) of all components, and analyse the statistical histogram of the three features. From the three histograms, it can be learned that, TCCs satisfy the defined verification criteria with $\xi_s = 0.7$, $\xi_c = 0.5$ and $\xi_a = 0.5$, while pictorial components are well eliminated.

In the proposed text segmentation algorithm, the complexity of the algorithm should be concerned. Assuming that there are total p pixels in one SCI, the LIAM-based coarse segmentation and the closing connection with the minimum scale (i.e., 3×3)

can be implemented in the time complexity of $o(p)$. The most complicated part is the proposed SOIG that needs to calculate the distances between each two characters. If the total number of remaining characters on the textual layer is q , it takes $o(q^2)$ times to compute the distances between each two characters. The proposed skeleton detection and separation algorithm can be implemented in $o(q)$. However, q is generally much smaller than p , as there are generally a few of characters with large sizes. In a summary, the segmentation is implemented by using non-optimized Matlab code on a regular PC(Intel(R) core(TM)2, 2.4 GHz, 4 GByte RAM). The average processing time for SCIs with the size 1280×1024 is about 0.5 s.

There are some cases that the proposed algorithm cannot settle well. As shown in Fig. 3.22, three sample images are displayed with text segmentation results using the proposed algorithm. In the upper image, more than two artistic texts with various scales and orientations are closely neighbored. Characters in each word are already connected, and the nearest distance between each two strings is very short (most of them are less than 3 pixels). Thus, after the image *colsing* operation with the minimum scale, all the texts are connected together, forming one connected component. This component may be classified as a pictorial component due to its failure to satisfy the defined TCC verification criteria. This situation also occurs in the right-bottom image. In the left-bottom image, some texts are connected with unstructured graphics. The finally formed components of these texts may fail to satisfy the TCC verification criteria due to their irregular shapes. These artistic designed images also cannot be well handled by existing methods, as described in [3, 33].

3.5 Summary

In this chapter, we have proposed a coarse-to-fine text segmentation framework for SCIs. In the coarse partitioning stage, the LIAM algorithm can highlight the difference between textual and pictorial regions and effectively eliminate most pictorial regions. The TCC-based refinement can locate textual regions precisely while maintaining the integrality of text characters, regardless of the variations of text scales and orientations. In particular, the proposed SOIG algorithm can adaptively connect adjacent characters to form the basic textual objects (i.e. TCCs) with uniform features, incorporating with the morphological operations. Additionally, over-connection is well handled to improve

Chapter 4

Learning Based Compression for Screen Content Images

4.1 Introduction

Transmission of computer screen images has been widely involved in many multi-device applications [1, 2, 5]. Generally, SCIs are transmitted on a low-bandwidth channel with the real-time requirement. Therefore, effective and efficient compression for SCIs is much desired. The screen image is a kind of compound image composed of textual and pictorial regions. The pictorial part can be compressed well using the traditional coding algorithms such as JPEG and JPEG2000, while these algorithms are inefficient to encode the textual part. In SCIs, textual parts are always with high-frequency components in strong and intensive edges. In addition, the difference between adjacent pixels in textual components are usually significant. Therefore, it is hard to remove the correlations in the text regions and achieve sparse representations by using the conventional basis functions (e.g. DCT or Wavelet transform). Compression schemes for textual components are then exploited based on the unique properties of textual regions, such as edge, color distribution, gradient, etc.

Therefore, the most advanced approach for SCI compression is segmentation-based, trying to decompose a screen image into different parts or layers, and then to process these parts separately [10, 12, 45, 46], as introduced in Chapter 2. However, some existing coding algorithms designed for natural images are still adopted for encoding textual regions, such as in [12], the Portable Network Graphics (PNG) coding algorithm

is employed to encode the textual layer. If specified basic functions are designed for textual regions, more sparse representation of texts may be achieved, such that the textual regions can be encoded more efficiently.

In this chapter, we propose a learning-based compression scheme for screen images. The compression diagram is shown in Fig. 4.1. The input screen image is firstly segmented into textual and pictorial blocks. A specialized text dictionary is learned to sparsely represent the textual blocks. Since there are typical characteristics in textual blocks, such as strong edges, non-continuous sample values, and limited distribution orientations in local areas, the learned text dictionary can represent these characteristics more effectively than the traditional basis functions. The K-SVD (Singular Value Decomposition) algorithm is adopted to train the text dictionary based upon a large set of various textual blocks. In particular, an over-complete separated version of DCT basis functions is designed to initialize the text dictionary to speed up the convergence. Based on the learned dictionary, the sparse coefficient matrix of textual blocks can be acquired and then be entropy-coded. As to the pictorial blocks and the segmentation index map, we utilize the conventional JPEG algorithm to encode them.

The remaining of this chapter is organized as follows. In Sec. 4.2, we introduce the dictionary learning method for text representation, and analyze the effectiveness of the resultant text dictionary. The proposed compression scheme is given in Sec. 4.3. Experimental results and conclusions are provided in Sec. 4.4 and Sec. 4.5, respectively.

4.2 Over-complete DCT based Text Dictionary Learning

In sparse signal representation, complete basis functions (e.g. DCT, Gabor and Wavelet) are generally employed to represent signals with sparse feature matrices. Most of these basis functions are designed for natural images and are inefficient to sparsely represent the screen images. During the past decade, over-complete dictionaries with fewer constraints, which are more flexible and adaptive to specific signals, has been investigated widely [107, 108, 109]. They can model the intricate data structures with much sparser coefficients compared with complete basis functions.

Figure 4.1: The proposed scheme for SCIs compression. The key steps in the scheme are highlighted by color filling in the figure.

4.2.1 Over-complete DCT based K-SVD Algorithm

Using an over-complete dictionary to sparsely represent a signal is commonly mentioned as sparse coding [110], which can be formulated as follows. Let $D = \{d_1, d_2, \dots, d_K\} \in \mathbb{R}^{N \times K}$ denote an over-complete dictionary, where the columns constitute the dictionary atoms and $K \gg N$. Using the over-complete dictionary D to represent a signal $y \in \mathbb{R}^N$ with a coefficient vector $x \in \mathbb{R}^K$, the sparse representation of the signal can be defined as follows.

$$x^* = \arg \min_x SC(x), \quad \text{subject to } y = Dx \quad (4.1)$$

where $SC(x)$ is sparse constraint of x (usually defined as l_0 norm to count the non-zero entries in a vector).

Here, we learn a text dictionary $D^T = \{d_1, d_2, \dots, d_K\} \in \mathbb{R}^{N \times K}$ based on a set of text patches $S = \{s_i\}_{i=1,2,\dots,n} \in \mathbb{R}^{N \times n}$, and meanwhile get a sparse representation $C = \{c_i\}_{i=1,2,\dots,n} \in \mathbb{R}^{K \times n}$ for S under the sparsity constraints. The learning process for

the text dictionary is formulated as an optimization process with variables D^T and C :

$$\begin{aligned} & \text{minimize } \|S - D^T C\|_2^2 \\ & \text{subject to } \forall i, \|c_i\|_0 \leq T_0 \text{ and } \forall j, \|d_j\|_2^2 = 1 \end{aligned} \quad (4.2)$$

where T_0 is the sparsity constraint and it is a predetermined maximum number of non-zero entries in each c_i .

The K-SVD algorithm [111, 112] can be used to handle the optimization process above. An over-complete dictionary of K atoms is learned based on the training set. The learning process is an iterating process including two stages: sparse coding and dictionary update. The sparse coding is completed by the pursuit algorithms such as Orthogonal Matching Pursuit (OMP), Basis Pursuit (BP) and FOCal Underdetermined System Solver (FOCUSS), while the dictionary update is implemented based on the Singular Value Decomposition (SVD) algorithm. In the K-SVD algorithm, it has been proved that the solution is an optimal one if T_0 is relatively small enough to N [112].

Although the K-SVD algorithm can guarantee the convergence due to the effectiveness of the pursuit algorithm, there are several critical factors affecting the convergence speed. An important one is the initial state of the dictionary. In the original K-SVD algorithm, the dictionary is initialized with a constant atom and $K - 1$ samples randomly selected from the training set. A large amount of iterations are required to get the convergence for the randomly selected samples due to their numerous variations. In order to speed up the convergence, we construct an over-complete separable version of DCT basis functions (denoted as over-complete DCT) as the initial atoms. They are obtained by sampling the cosine wave in different $\sqrt{K}$ frequencies in horizontal and vertical directions. The first column of the dictionary is set to a constant to extract the mean value of all blocks, denoted as the DC atom. The over-complete DCT basis functions are defined as:

$$\begin{aligned} F_{uv} &= \frac{1}{4} C_u C_v \sum_{x=0}^{m-1} \sum_{y=0}^{m-1} \cos(u\pi \frac{2x+1}{2m}) \cos(v\pi \frac{2y+1}{2m}) \\ & \text{for } u, v = 0, 1, 2, \dots, \sqrt{K} - 1. \end{aligned} \quad (4.3)$$

where

$$C_u/C_v = \begin{cases} 1/\sqrt{2} & \text{if } u/v = 0 \\ 1 & \text{otherwise} \end{cases}$$

Table 4.1: Convergence Comparison Result of overcomplete DCT based Learning and random sample based learning.

Iteration Number	Overcomplete DCT based			Random Sample based		
	MSE	NoC	Runtime (h)	MSE	NoC	Runtime (h)
10	0.0052	15.63	0.41	0.284	16.72	0.49
20	0.0048	15.14	0.67	0.2592	16.42	0.81
40	0.0029	13.62	1.12	0.2093	16.15	1.52
60	0.0016	12.69	1.86	0.1831	15.86	2.14
80	0.001	10.84	2.72	0.176	15.39	2.83

Based upon the over-complete DCT of K columns, a more adaptive dictionary can be obtained through the learning process. Our experiments show that the over-complete DCT based K-SVD algorithm can achieve more time-efficient convergence than the random sample based algorithm (as can be seen in Table 4.1).

The other one is the number of the dictionary atoms, K . In the K-SVD algorithm, K should be much larger than the dimension of each sample N . Generally, given T_0 , we can obtain better reconstruction performance with greater K . But, it will cost more time in training. Furthermore, much more bits have to be used to code the corresponding coefficient matrix. Hence, there is a trade off between the representation performance and the coding cost for determining the value of K . We are to verify the impact of K on the text representation in the following section.

4.2.2 Text Dictionary Learning

To learn the over-complete dictionary D^r for text signals by the K-SVD algorithm, we have to construct the training samples. Our training set includes two types of images: one is the textual images collected from various web pages and text documents, while the other one includes the textual images composed of isolated machine-printed characters (comprise 10 digits, lowercase and uppercase letters). These letters are with different typefaces, styles orientations and sizes. Then, we adopt an 8×8 window to scan all the images without overlapping, thereby acquiring numerous text patches. We discard all the blank text patches, and randomly select 36800 patches as the training set. All the 8×8 patches are reshaped to 64×1 vectors to form the training vectors.

Figure 4.2: Over-complete $DCT441$ (left) and $D^\tau 441$ (right).

Using the generated training vectors, we learn the text dictionary by using the proposed over-complete DCT based K-SVD algorithm. Just as in [111], K is set to 441 here. Experimental results show that it is a suitable value to ensure the reconstruction performance and the sparsity constraint. In the learning process, all the training vectors are normalized to $[0, 1]$. The maximum number of iterations is set to 80 and the reconstruction error ε is set to 0.001. Setting the maximum number of non-zero coefficients T_0 to 20, the Mean Squared Error (MSE) of approximation, as well as the average number of coefficients (NoC) for each sample, can be computed. If the representation error is less than ε or the number of iterations is greater than 80, the iterations will stop. From TABLE 4.1, we can see that the over-complete DCT based K-SVD algorithm can get the convergence faster.

The learned text dictionary consists of 441 normalized atoms in 64 dimensions and it is denoted as $D^\tau 441$ (64×441). As the over-complete dictionary is fixed, the sparse representation of any textual blocks can be calculated by pursuit algorithms. In this work, we adopt the OMP algorithm due to its simplicity and efficiency [113]. The original textual blocks can be reconstructed with high precision by multiplying the dictionary D^τ with related sparse coefficient matrix C . Fig. 4.2 illustrates the over-complete $DCT441$ and the learned dictionary $D^\tau 441$. This figure shows that the atoms in $D^\tau 441$ are more adaptive to text characters than those from the over-complete $DCT441$.

4.2.3 Sparsity Analysis of Text Representation

In order to verify the efficiency of the learned dictionary D^T441 , we randomly select 6000 textual blocks with the size of 8×8 from the web pages and text documents (different from the training set). We also learn several dictionaries for different K with the same training set and the same parameters, including D^T64 (64×64) and D^T256 (64×256). Experiments show that these specialized dictionaries can represent the text better (leading to sparser coefficients) than the DCT. In order to test the sparsity, we compare those dictionaries by using D^T441 , D^T256 , D^T64 , over-complete $DCT441$ and traditional DCT basis $DCT64$ to represent the testing set respectively.

We test the reconstruction performance in terms of PSNR and the number of non-zero coefficients for each block. For the i -th time, we pick the first i maximal coefficients of each block using the OMP algorithm under the objective function (4.2), and calculate the PSNR values of the reconstructed patches. Results are shown in Fig. 4.3.

Figure 4.3: Reconstruction results based on different dictionaries.

From Fig 4.3, we can see that the performance of D^T441 is much better than those of D^T64 and D^T256 , which is consistent with our analysis on the number of dictionary atoms in Sec.II (A). We can also see that the reconstruction performance of D^T441 is

Figure 4.4: Energy compaction of non-zero coefficients.

better than $DCT64$ and over-complete $DCT441$ based on the same number of non-zero coefficients. In particular, the PSNR of D^T_{441} is at least 5 dB higher than that from $DCT441$, and at least 8 dB higher than that from $DCT64$, when there are more than 6 non-zero coefficients. The reason is that the learned dictionary is more adaptive to textual components. Fewer dictionary atoms are needed to reconstruct the original samples. The experimental results show that the learned dictionary can effectively represent textual images regardless of the specific text characters, e.g., typeface and size. More importantly, compared with fixed DCT, the text representation generated from the learned dictionary is much sparser and this guarantees better compression performance than DCT-based coding algorithms.

The energy compaction of acquired coefficients is also analyzed here, as shown in Fig. 4.4, since the energy compaction can reflect the contribution of coefficients. The largest coefficient represents the most significant component in a block. For a coefficient vector $c \in \mathbb{R}^K$, the energy ratio of the first k maximal coefficient values for each block is defined as

$$R_{energy} = \sum_{j=1}^k c(j)^2 / \sum_{j=1}^K c(j)^2 \quad (4.4)$$

We analysis the energy ratios of the first $k \in \{1, 2, \dots, 20\}$ maximal coefficients, using

different dictionaries. Fig. 4.4 shows that $D^{\tau 441}$ achieves the best energy compaction.

4.3 The Proposed SCI Compression Scheme

In this section, we describe our proposed compression scheme as illustrated in Fig. 4.1. The proposed algorithm consists of three main procedures: dictionary learning, image segmentation and compression. The dictionary training process is an off-line procedure and precedes the image compression procedure, which has been described in Sec. 4.2. We will present the detailed compression process in the following subsections.

4.3.1 Preprocessing of SCIs

It is obvious that there are more sharp edges in the textual blocks than those in the pictorial blocks. In other words, textual blocks have more high-frequency components. When compressing these two types of blocks by the traditional coding algorithms (e.g. JPEG, SPIHT[114] and JPEG2000) at a given compression ratio, the PSNR values vary greatly. Two natural images (*lena*, *barbara*) and two general textual images (*text1*, *text2*) with the size of 512×512 are encoded to help the analysis. The compression results of some natural and textual images in terms of PSNR values are shown in Table 4.2.

From Table 4.2, we can see that the PSNR values for the textual images are much lower than those for natural images at the same compression ratio. This demonstrates the inefficiency of the traditional coding algorithms (JPEG, SPIHT, and JPEG2000) for compressing the textual images. Therefore, segmentation-based coding methods are proposed for SCI coding. In this work, we adopt the TCC-based segmentation algorithm introduced in Chapter 3 to decompose the screen images into textual and pictorial blocks.

A segmentation index map and two categories of blocks are generated from this algorithm. We then encode these two types of blocks separately. The learned text dictionary also includes some low-frequency dictionary items, which have the capability to approximate high-frequency pictorial blocks. Based upon the learned dictionary, the sparse coding can achieve acceptable coding performance for natural images, as can be seen in Table 4.2.

Table 4.2: PSNR values (dB) for test images at given bitrates.

Algorithms:	JPEG		SPIHT		JPEG2000		Sparse Coding	
Bitrate(bpp):	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0
lena	35.0	37.7	37.2	40.5	37.2	40.3	34.2	37.4
barbara	28.3	33.2	31.4	36.4	32.2	37.1	27.2	31.9
text1	16.0	19.3	17.7	22.8	18.6	24.1	23.1	28.4
text2	16.7	19.8	18.0	22.8	18.7	23.9	24.2	28.8

4.3.2 Coding of Sparse Textual Representation Matrix

There are three kinds of components to be coded after segmentation: textual blocks, pictorial blocks and segmentation index map. As to the textual blocks, the sparse coefficient matrix $C \in \mathbb{R}^{441 \times n}$ can be calculated by using the OMP pursuit algorithm based on the learned text dictionary D^{T441} (sparse coding). If the representation error is less than ε or the number of non-zero coefficients of each block is larger than T_0 , the iterations will be terminated.

Since the first DC atom of the dictionary is fixed, the first value of each column in C is much larger than the other values in the same column. Here, the Differential Pulse-Code Modulation (DPCM) encoder is applied to code the first row of C . As the rest of the coefficient matrix (C') usually has less than T_0 non-zero coefficients in a vector of K dimensions ($K \gg T_0$), a position-based coding method is proposed. The positions and the values of the non-zero coefficients in C' are recorded. As to the row positions, from bottom to the top of each column, a subtraction operator is applied to the neighboring values to reduce the redundance. A simple example of this coding method is shown in Fig. 4.5. The left one shows the non-zero coefficients in C' , while the right figure records the number and the row positions of non-zero coefficients in each column. The number of non-zero coefficients in each column, the new positions and values are then entropy-coded.

The conventional JPEG coding method is used to code the pictorial parts and the JBIG is adopted to encode the index map. Note that the text dictionary does not need to be transmitted, as it is supposed to be known in both encoder and decoder.

	1	2	3	4	5	...
1					20	
2		21	18			
3	24			12	10	16
4						
5		12	9			
6						
7				3		4
8	9	2				

Num. of Non-zero in each column	Row Position	New Position	Values
2	3	3	24
	8	5	9
3	2	2	21
	5	3	12
	8	3	2
	2	2	18
	...	...	...

(a) Sparse coefficient matrix C'

(b) information of non-zero coefficients

Figure 4.5: Position-based coding for the sparse coefficient matrix C' .

4.4 Experimental Results

In our experiments, SCIs with the size of 1280×1024 are used to demonstrate the performance of the proposed scheme. They include webpages, PDF files and combined images displayed on compute screens. Four typical images are selected to illustrate the performance, as shown in Fig. 4.6 (a)-(d) respectively: webpage with binary text, webpage with colored text, Power Point image and a combined image with pictorial and textual parts. Percentage of textual regions in these SCIs is different. Those images are scanned by a 8×8 window, and segmented into textual blocks and pictorial blocks by using the TCC-based segmentation algorithm. They are then encoded by the proposed compression scheme.

In order to verify the efficiency of the learning based compression algorithm, we compare the coding performance from the proposed method with that from the traditional compression methods (e.g. JPEG, JPEG2000). We compare the visual quality of the reconstructed images with the same compression ratios. Fig. 4.7 shows the comparison results of a sub-image cropped from image webpage2 at the bit-rate of 1.0 bpp. The visual quality of the reconstructed images from the proposed method outperforms those from others greatly. The reconstructed images from the JPEG have annoying blocking artifacts around the edges of the text characters, as shown in Fig. 4.7(b). The

(a)

(b)

(c)

(d)

Figure 4.6: Compression results for four example SCIs: (a) is a webpage with binary texts; (b) is a webpage with colored texts; a slide image and a combined screen image are shown in (c) and (d).

Figure 4.7: Quality comparison of reconstructed images (bitrate:1.0 bpp).

reconstructed images from the JPEG-2000 become blurring, as shown in Fig. 4.7(c).

Based on the proposed compression scheme, the R-D curves of the luminance part for the four images are shown in Fig. 4.8 (a)-(d). From this figure, we can observe that the proposed scheme with the over-complete dictionary $D^{\tau 441}$ outperforms the JPEG and JPEG2000 algorithms for all images. The PSNR values from $D^{\tau 441}$ are 10 dB higher than those from the JPEG algorithm on average. Note that Zig-Zag scanning method and EOB coding are not suitable for the sparse coefficient matrices because of the irregular distribution of non-zero coefficients. Although the position-based coding introduces extra bits to code the position information, the results is still satisfactory due to the advantage of the sparse text representation over DCT basis functions.

Additionally, four dictionaries ($D^{\tau 441}$, $D^{\tau 256}$, $D^{\tau 64}$ and $DCT441$) are employed to test the coding performance. From Fig. 4.8, it is shown that the PSNR of the results from $D^{\tau 441}$ is much better than those from $D^{\tau 256}$ and $D^{\tau 64}$. As we know, a larger

Figure 4.8: RD curves of Compression results for four screen images.

value of K results in much better reconstruction performance, which can compensate the cost in encoding the irregular coefficient matrix. Moreover, the result from D^T_{441} is much better than that from DCT_{441} , which confirms the effectiveness of the proposed learning based algorithm.

The segmentation process can be implemented with the cost of $O(P)$, where P is the pixel number in the image. To assess the computational complexity of the sparse coding, we discuss the encoding and the decoding processes separately. The encoding process is implemented by the OMP algorithm, whose computational complexity is $O(bKN_a)$ per patch, where b is the size of the patch, K is the number of atoms in the dictionary,

and N_a is the average number of non-zero values in the coefficient vector. Therefore the computational complexity of the entire encoding process is $O(\frac{P_t}{b}bKN_a) = O(P_tKN_a)$, where P_t is the number of pixels in all the textual blocks. The decoding stage is simply implemented by the multiplication of the dictionary and related coefficients. The operation can be done with the cost of $O(bN_a)$ per patch. The overall complexity of this stage is therefore $O(\frac{P_t}{b}bN_a) = O(P_tN_a)$.

The segmentation and compression has been all implemented using non-optimized Matlab code on a regular PC (Intel(R) core(TM)2, 2.4 GHz, 4 GByte RAM). The compression of screen images with the size 1280×1024 requires 1.2 s on average, and the decompression takes less than 0.3 s.

4.5 Summary

In this chapter, we have proposed a learning based computer screen image compression method, to address the increasing need of handling such images. Firstly, we have performed careful studies on the characteristics of screen images as this is still a less-investigated area so far. A specialized over-complete dictionary for textual blocks is then learned. The learned dictionary can represent the text more sparsely and efficiently compared with the traditional DCT basis functions. Based on our proposed compression scheme, the computer screen images can be transmitted at a low bit-rate with high quality of the reconstructed images. In our future work, we are on the way of solving the sparse coding problem with more constraints to obtain a more adaptive dictionary for screen image compression.

Chapter 5

Perceptual Quality Assessment of Screen Content Images

In this chapter, we aim to carry out the first in-depth study on perceptual quality assessment of SCIs. We present the study on visual quality assessment of distorted SCIs from both subjective and objective aspects. In the subjective evaluation, a large-scale *Screen Image Quality Assessment Database (SIQAD)* is constructed with corresponding *Difference of Mean Opinion Scores (DMOS)* values. For the objective assessment, a new quality assessment metric, i.e. *SCI Perceptual Quality Assessment (SPQA)*, is proposed, which has high consistency with HVS as viewing SCIs. To the best of our knowledge, the *SIQAD* is the first large-scale database published for quality evaluation of SCIs, and this is the first attempt to explore the perceptual quality assessment of distorted SCIs.

5.1 Introduction

Research on quality assessment of SCIs becomes important as it is playing important roles in many SCI processing procedures. As introduced in Chapter 1, when processing SCIs, various distortions may be involved, such as noise, blurring, contrast change and compression artifacts. Firstly, in order to compare the performance of different processing methods, an uniform metric should be adopted to evaluate the final quality of processed images, and this metric should have high consistency with HVS as human beings are the ultimate viewers of the images. Secondly, the quality assessment met-

rics can be utilized to guide the processing procedures, for example, optimize the bit rates and visual quality simultaneously in the SCI compression, such that the finally generated images will have good visual quality, highly consisting with HVS. Therefore, perceptual quality assessment of SCIs is much desired for these processing methods. Although many IQA methods have been proposed to evaluate quality of natural images [30], whether these IQA methods can be applicable to evaluate visual quality of SCIs is still an open question.

In this work, we firstly construct a large-scale image database for the subjective testing, i.e. *SIQAD*, consisting of 20 source and 980 distorted SCIs. The source SCIs have various content styles, cutting from webpages, slice files, digital magazines and pdf files. The distorted SCIs are derived from seven different processes: Gaussian noise (GN), Gaussian Blurring (GB), Motion Blurring (MB), Contrast Change (CC), JPEG, JPEG2000 and Layer Segmentation based Compression (LSC), and each is at seven degradation levels. In order to get the subjective quality scores and investigate which part (text or picture) contributes more to the overall visual quality, the 11-category *Absolute Category Rating* (ACR) [62] is employed to obtain three subjective scores corresponding to the entire, textual and pictorial regions, respectively. The correlation between these three groups of scores can then be analysed to reveal the contribution of textual or pictorial regions.

Based on the results of the subjective testing, we propose a new scheme, *SCI Perceptual Quality Assessment* (SPQA), to objectively evaluate the quality of distorted SCIs. The SPQA consists of an objective metric and a weighting model. The objective metric is designed to separately evaluate the quality of textual and pictorial regions. In particular, a new integration scheme is designed to combine the luminance and sharpness similarity between reference and distorted SCIs in the objective metric, considering the different characteristics of textual and pictorial regions. The weighting model, which is constructed based on the analysis of the subjective scores, combines the predicted quality scores of textual and pictorial regions to the overall quality of tested SCIs.

Finally, to investigate the applicability of existing objective IQA metrics, 11 advanced IQA approaches are employed to evaluate the quality of SCIs in *SIQAD*. Through detailed analysis, we found that existing IQA methods are limited in predicting the quality of the distorted SCIs. The proposed SPQA scheme achieves much higher consistency

with human visual perception when judging the quality of distorted SCIs, compared with the 11 IQA methods.

In the remaining of this chapter, we introduce the configuration of the *SIQAD* and the subjective assessment methodology in Sec. 5.2. The proposed objective metric for SCIs as well as the weighting model is presented in Sec. 5.3. In Sec. 5.4, we report the experimental results, followed by summary drawn in Sec. 5.5.

5.2 Construction of the Screen Image Quality Assessment Database (SIQAD)

To investigate quality evaluation of SCIs, we construct a large-scale screen image database (i.e. *SIQAD*) with seven distortion types, each with seven degradation levels. Totally, 20 reference and 980 distorted SCIs are included in the *SIQAD*. Subjective test of these SCIs is conducted to obtain the subjective quality scores. All the SCIs and the subjective scores are available on the website [115].

5.2.1 Generation of Distorted SCIs in the *SIQAD*

The reference SCIs are selected with various layout styles, including different sizes, positions and ways of textual/pictorial region combination. Meanwhile, pictorial or textual regions are also diverse in visual content. The percentage of textual regions on these SCIs are also varied. In total, twenty SCIs are collected from webpages, slides, PDF files and digital magazines through screen snapshot. The reference SCIs are cropped from these twenty images to proper sizes for natively displaying on computer screens during the subjective test.

Seven distortion types which usually appear on SCIs are applied to generate distorted images. Gaussian noise is often involved in image acquisition, and is thus included in most existing image quality databases [56, 58]. Gaussian blur and motion blur are also considered due to their commonly existing in practical applications. For example, when capturing SCIs using digital cameras, hand-shaking, out-of-focus or object moving would bring blur into images. Contrast change is also an important factor affecting peculiarities of the HVS. Different settings of brightness and contrast of screens will result in various visual experiences of viewers. As compression of SCIs

is widely used processing in most multimedia applications, three commonly used compression algorithms are utilized to encode the reference SCIs: JPEG, JPEG2000 and *Layer Segmentation based Coding* (LSC) [11]. The JPEG and JPEG2000 are two widely used methods to encode images, and have been introduced into many quality assessment databases. We include LSC as another codec due to its efficient compression of SCIs. The LSC firstly separates SCIs into textual and pictorial layers, and then encodes the textual layers by the *Basic Colors and Index Map* (BCIM) method [11] while the pictorial layers are encoded by the JPEG algorithm.

For each distortion type, seven levels are set to generate images from low to high degradation levels. The detailed configuration of these algorithms is given as follows.

- Gaussian Noise (GN): distorted images are generated by the MATLAB function `imnoise` with zero mean. The standard deviation is set from 0.02 to 0.24: 0.02, 0.04, 0.06, 0.1, 0.14, 0.18 and 0.24.
- Gaussian Blurring (GB): blurred images are obtained by implementing the MATLAB function `imfilter` with Gaussian kernel. Size of the Gaussian kernel is 7 by 7, and the standard deviation is set to 0.58, 0.68, 0.76, 0.96, 1.2, 1.8 and 2.4.
- Motion Blurring (MB): motion blur is added to images by using the MATLAB function `imfilter` with the motion kernel that approximates the linear motion of a camera. Two parameters `len` and `theta` are set to control the blurring level.
`Len`: linear motion of a camera by `len` pixels;
`Theta`: an angle degree in a counterclockwise direction.
 Here, `len` is set to [3, 3.4, 3.6, 4, 4.8, 6, 9], and `theta` is zero.
- Contrast Change (CC): MATLAB function `imadjust` is used to change contrast of test images. The contrast changes are listed as follows. Left is the scale of original image, and right is the scale of contrast changed image. [0 1] to [0.3, 0.5], [0 1] to [0.1, 0.7], [0.1, 0.8] to [0.1, 0.9], [0.2, 0.8] to [0.1, 0.8], [0.1, 0.6] to [0.1, 0.8], [0.2, 0.6] to [0.1, 0.8], [0.2, 0.7] to [0 1].
- JPEG compression: the JPEG is implemented via MATLABs `imwrite` function. The quality factor is set to [75, 55, 48, 32, 25, 18, 13].

- JPEG2000 compression: MATLAB function `imwrite` is also used to generated images compressed by JPEG2000. The compression factor is set to [20, 28, 32, 40, 46, 54, 65].
- Layer-Segmentation based Compression (LSC): we change the classification index map artificially. Some textual blocks are randomly misclassified to pictorial regions with misclassification ratio ranging from 40% to 80%. Quantization factor is set to [10, 8, 7, 6, 5, 4, 3] for the textual layer, while the pictorial layer is gradually of better quality (with the quality factor of JPEG equally spaced from 15 to 60).

These seven level distortions can create a broad range of image impairment types, such as blurring, blocking, structured distortion and misclassification artifacts. Specifically, in the LSC method, misclassification is added to compressed SCIs by adjusting the segmentation index map artificially. Some textual blocks are randomly misclassified to pictorial regions with different misclassification ratios. Some distorted SCIs are shown in Fig. 5.1 for readers' observation. The distorted images are labeled as $cimd_p_q$, where $d \in \{1, 2, \dots, 20\}$ is the index of the reference SCIs; $p \in \{1, 2, \dots, 7\}$ indicates the distortion types (e.g. GN:1, GB:2, MB:3, CC:4, JPEG:5, JPEG2000:6 and LSC:7); $q \in \{1, 2, \dots, 7\}$ denotes the degradation level.

5.2.2 Equipment Configuration and Human Subjects

According to the recommended viewing conditions in laboratory environment [62], the experiments are conducted using identical desktop with 16 GB RAM and 64-bit Windows operating system. The desktop is placed in a laboratory room with normal indoor light. Calibrated LED monitors (Dell P2412H) of 24 inches are equipped to display images in *SIQAD* at a resolution of 1920×1080 . All subjects are required to sit at an approximate viewing distance of 2 – 2.5 times of screen heights. The graphical user interface is shown in Fig. 5.2. All SCIs can be displayed at their native resolutions to prevent any distortions from scaling operations. Subjects report their judgements of by clicking a option button, and values of the option buttons are automatically reset after each evaluation.

Figure 5.1: Examples of distorted SCIs. Due to the space limitation, only one level distortion is illustrated (refer to the database for more images).

Michelle Whybrow, New Zealand Birds and Mammals, Auckland Zoo

In September 2011, Auckland Zoo opened Te Wao Nui, a \$16m New Zealand precinct, which takes up one-fifth of the Zoo's 17 ha footprint and offers visitors a unique experience of this country's native animals, plants and culture.

The new development is home to around 60 different New Zealand native species and encompasses six distinct ecological habitats – The Coast (Takurua), The Islands (Moutere Rahui), The Wetlands (Nga Repu), The Night (Te Poi), The Forest (Te Wao Nui a Tane), and The High Country (Whenua Waotui).

Our four male kea share a 520m² by 9 metre high aviary with weka within The High Country zone, which replicates the South Island's isolated high country environment with schist rock, river stone, tussock grasses and a tramper's hut. It is also home to blue duck (whio), kakariki, Ottago skink and rough gecko.

retaining wall, stabilising a steep cut in the existing ground on the side of the hill. This provides the maximum range of topography within the aviary, allowing great range of movement for both the kea and weka. The rock wall has been

water through the aviary and helps keep the water oxygenated. Special areas have also been created in the stream to allow all the birds easy access to bathe, while allowing the visitors opportunity to enjoy observing these situations.

Subjective Testing of Digital Compound Images

Please give a score in [0,10] to evaluate quality of the image:

Bad: very annoying;
 Poor: annoying;
 Fair: slightly annoying;
 Good: perceptible, but not annoying;
 Excellent: imperceptible

BadPoorFairGoodExcellent

012345678910

Image ID: total images:

Please evaluate the quality of the specialized regions based on your visual perception:

---for the entire image

Scores

01234

5678910

Previous
Next

Figure 5.2: Graphical user interface in the subjective test. The red tooltip will change if subjects need to judge different regions.

The participants taking part in the study are convened from Nanyang Technological University, Singapore. Total 96 subjects participate in the subjective testing. The subjects are all university under-graduate or graduate students, aging from 19 to 38 years old. Most of them are students inexperienced with image visual quality assessment, and their vision are confirmed to be normal or corrected. Before the subjective testing, each subject is individually briefed the way to do the image quality judgement and given a short training to get familiar to the operations in the testing procedure.

5.2.3 Subjective Testing Methodology: Absolute Category Rating (ACR) for Assessing Quality of SCIs

Subjective testing methodologies of image quality evaluation have been recommended by ITU-R BT.500-13 [62], including *Absolute Category Rating* (ACR), double-stimulus impairment scale and paired comparison. In this study, 11-category ACR is employed. Given one image displaying on the screen, the human subject is asked to give one score (from 0 to 10: 0 is the worst, and 10 is the best) on the image quality based on her/his visual perception. The quality scales are briefly introduced to subjects, as shown in the left corner in Fig. 5.2. This methodology is chosen because the viewing experience

of subjects is close to that in practice, where there is no access to the reference images. We report the details of the subjective experiments as follows.

In this study, we would like to not only get the visual scores of all distorted SCIs, but also investigate which part (text or picture) contributes more to the overall visual quality. Hence, subjects were required to give three scores to each test image, corresponding to overall, textual and pictorial regions, respectively. In this test, each image was shown three times inconsecutively, and subjects gave one score to one specific region at a time. A highlighted prompting box (i.e. the red text string as shown in Fig. 5.2) is designed to remind subjects which part is being judged. When judging one image, three aspects are mainly considered: content recognizability, clarity and viewing comfortability. All the reference images are also included in the test. We generate a random permutation of 1,000 images (20 references and 980 distorted SCIs) for each round, and make sure that every two consecutive images are not generated from the same reference image. According to [62], the execution time of one test session should not exceed 30 minutes to avoid user fatigue. Thus, we split each permutation into 8 groups and assign one group of images to one subject at a time. Each subject can finish the evaluation of several groups. Totally, 96 subjects took part in the study, and each image was evaluated by at least 30 subjects.

5.3 Perceptual Quality Assessment Of SCIs

As aforementioned in Chapter 2, due to the different properties of textual and pictorial regions in SCIs, the same distortion in different regions may lead to different visual perception of human beings. Hence, it is natural and reasonable to separately handle each part, and then combine them together with differentiation. In this section, we firstly investigate which part (text or picture) attracts more visual attention as viewing the distorted SCIs. Based on the results, an activity-based weighting model is proposed to combine the two parts to predict the overall quality of the SCIs. Then, we propose a novel quality assessment metric to objectively measure the quality of textual and pictorial regions, considering their visual differences. The diagram of the proposed *SCI Perceptual Quality Assessment* (SPQA) method is illustrated in Fig. 5.3. One reference and its distorted version are firstly segmented into textual and pictorial layers based on the text segmentation algorithm proposed in [95]. The quality of the textual and

pictorial layers of the distorted SCI are separately evaluated by the proposed objective metric. The two quality measures are then combined using the proposed weighting model, to get the final visual quality score of the distorted SCI.

Figure 5.3: Diagram of the proposed algorithm SPQA. The SPQA mainly contains two algorithms highlighted in the figure.

5.3.1 Which Part Attracts More Visual Attention: Textual or Pictorial?

After getting the subjective scores (QE , QT and QP) of entire, textual and pictorial regions respectively, one problem we would like to explore is the correlation among these three scores. Through in-depth investigation of this correlation, an effective way for

integrating textual and pictorial parts can be figured out. Here we make some initial investigations on the combination of QT and QP , based on which a new weighting model is proposed to estimate the final overall quality of SCIs.

There are many factors affecting human perception when viewing SCIs, including area ratio and position of texts, size of characters, and content of pictures, etc. In the proposed model, we investigate a statistical property of SCIs that reflects impairments of test images, rather than any specific factor. Image activity reflects the variation of image content, which can be used to differentiate images [116, 117]. Based on the activity measure and the segmentation algorithm proposed in [95], we propose a novel model to compute two weights ($\mathcal{W}_t$ and $\mathcal{W}_p$) that can measure the effect of textual and pictorial regions to the quality of the entire image. In particular, given one reference SCI and its text segmentation index map T , we calculate the activity map A of the corresponding distorted SCI [95]. The activity maps A_t and A_p of the textual and pictorial regions can be calculated based on A and T . Considering the human visual activity in the HVS (points closed to the fixation center are with high visual acuity, while other points far away are with relative low visual acuity), a Gaussian mask G is used to weight the activity values. Based on the weighted activity map, we compute two activity values for the textual and pictorial parts respectively, which are subsequently employed as weights to combine QT and QP .

The weighting model is constructed as a linear combination of QT and QP as follows.

$$QE' = \mathcal{W}_t * QT + \mathcal{W}_p * QP \quad (5.1)$$

where

$$\mathcal{W}_t = \frac{\sum_{i=1}^m \sum_{j=1}^n (A * T * G)_{i,j}}{\sum_{i=1}^m \sum_{j=1}^n (T)_{i,j}} \quad (5.2)$$

and

$$\mathcal{W}_p = \frac{\sum_{i=1}^m \sum_{j=1}^n (A * (1 - T) * G)_{i,j}}{\sum_{i=1}^m \sum_{j=1}^n (1 - T)_{i,j}} \quad (5.3)$$

are weights for textual and pictorial regions, respectively. Following the same notation above, m and n represent the dimensions of the test image. QE' is the predicted quality score of the distorted SCI. The performance of the proposed model will be assessed by calculating the correlation between the predicted score QE' and the subjective score QE .

5.3.2 Objective Quality Assessment of SCIs

It is known that the HVS is highly sensitive to the following properties of images: luminance, contrast and sharpness. These properties change along with various image distortions, such as noise corruption, blur, quantization and compression artifacts. Hence, they have been widely investigated in the FR NIQA. In SSIM [68], the product of three components of similarity is computed to estimate the image local quality at the location (x, y) :

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma \quad (5.4)$$

where $l(x, y)$, $c(x, y)$ and $s(x, y)$ are luminance, contrast and structural similarity; α , β and γ are positive constants used to adjust the relative importance of these three components. A simple setting ($\alpha = \beta = \gamma = 1$) is adopted in SSIM and most of its variations [68]. Liu et al.[75] used gradient similarity to replace the contrast/structural similarity in SSIM, and proposed a weighting method to combine the luminance and gradient similarity as follows:

$$q = (1 - W) \times g(x, y) + W \times e(x, y) \quad (5.5)$$

where q is the quality score of a local patch at the location (x, y) ; $e(x, y)$ and $g(x, y)$ are luminance and gradient similarity. $W = 0.1 \times g(x, y)$ is used as weighting values to highlight the contribution of the gradient similarity to the final quality. In [76], the authors found that, without any additional information, using the image gradient magnitude alone can yield highly accurate quality prediction. It is not only effective in estimating the visual quality, but also efficient to be implemented in real time applications.

However, these interaction schemes of the properties cannot work well for SIQA, since human visual perception to textual and pictorial regions are different. The distortions in textual regions are not always playing the same role to the overall quality. For example, subjects can easily notice luminance and contrast change in pictorial regions. However, they prefer to give high quality scores to texts with high integrity and clear shape, even though their color intensity or contrast has been greatly changed. Conversely, subjects are more sensitive to blurring artifacts appearing on textual regions than that on pictorial regions. As illustrated in Fig. 5.4, there is motion blur appearing on the image in (b) and color intensity change occurring on the image in (c). We can

(a) Reference image: cim11

(b) cim11.3.5, DMOS:63.98

(c) cim11.4.7, DMOS:37.50

(d) cim11.4.1, DMOS:76.54

Figure 5.4: DMOS values of some examples in *SIQAD*. The scale of the DMOS values is from 0 to 100. A higher value represents worse visual quality of the image (refer to the images at the original resolution for better visual comparison).

see that the background content and color intensity of texts in (c) are much different from the reference image in (a), while the background and contrast of texts in (b) are well maintained. However, subjective test shows that humans are more satisfied with (c) than (b), which can be reflected from their DMOS values: 63.98 for (b) and 37.50 for (c). Therefore, in these cases, we should reduce the effect of the luminance change to the overall quality of textual regions. However, with much luminance change, as displayed in Fig. 5.4 (d), subjects give low quality scores to this image at their first impression. Hence, for these cases, the effect of the luminance change in textual regions to the overall quality should be enhanced.

Based on the above analysis, we propose a new scheme for quality evaluation of distorted SCIs. In the proposed scheme, sharpness and luminance similarity between the reference and distorted SCIs is computed. Sharpness is computed since it is a good measure to summarize various distortions appearing in images [118]. The luminance similarity of textual regions is adaptively integrated to the sharpness similarity, while only sharpness similarity is considered for pictorial regions. For one SCI X and its distorted version Y , given its text segmentation index map T in which textual pixels are marked by one and pictorial pixels by zero, their textual layers (X_t, Y_t) and pictorial layers (X_p, Y_p) are calculated by $X_t = X \cdot T$, $X_p = X \cdot (1 - T)$, $Y_t = Y \cdot T$ and $Y_p = Y \cdot (1 - T)$. The luminance similarity map $S_l(X_t, Y_t)$ between the textual layers X_t and Y_t is calculated as follows:

$$S_l(i, j) = \frac{2 \cdot \mu_{xt} \cdot \mu_{yt} + c_1}{\mu_{xt}^2 + \mu_{yt}^2 + c_1} \quad (5.6)$$

where μ_{xt} and μ_{yt} denote the local mean values for each pixel (i, j) in the textual layers X_t and Y_t . c_1 is a parameter to avoid the instability when the denominator is close to zero.

To compute the sharpness of images, we use the multi-directional filters $\{h_k\}_{k=1,2,3,4}$ illustrated in Fig. 5.5. These filters can capture the local variations of images at four directions, including horizontal and vertical directions. The sharpness of one image X is measured by the summary of the first two maximum filtering results:

$$s(X) = |X \cdot h_a| + |X \cdot h_b|; \quad (5.7)$$

where a and b are the index of the filter that lead to the first two maximum results; $|\cdot|$ represents the absolute value of the convolution of X and h_k . Thus, the sharpness

0	0	0	0	0	0	0	1	0	0	0	0	1	0	-1	0				
1	3	8	3	1	0	8	3	0	0	0	0	3	8	-3	0				
0	0	0	0	0	1	3	0	-3	-1	-1	-3	0	3	-8	0				
-1	-3	-8	-3	-1	0	0	-3	-8	0	0	-8	-3	0	-3	0				
0	0	0	0	0	0	0	-1	0	0	0	0	-1	0	-1	0				
h_1					h_2					h_3					h_4				

Figure 5.5: Filters for calculating the sharpness values.

similarity between X_t and Y_t , and X_p and Y_p , are computed as:

$$S_s^t(i, j) = \frac{2 \cdot s(X_t) \cdot s(Y_t) + c_2}{s(X_t)^2 + s(Y_t)^2 + c_2} \quad (5.8)$$

$$S_s^p(i, j) = \frac{2 \cdot s(X_p) \cdot s(Y_p) + c_2}{s(X_p)^2 + s(Y_p)^2 + c_2} \quad (5.9)$$

where c_2 is a parameter to avoid the instability when the denominator is close to zero.

The quality map Q_map can be calculated by integrating the luminance and sharpness similarity maps as follows:

$$Q^\lambda_map(i, j) = [S_l]^\alpha \cdot [\lambda \cdot S_s^t + (1 - \lambda) \cdot S_s^p]^\beta \quad (5.10)$$

where $\lambda = \{0, 1\}$ is used as a flag to indicate which part is being processed: $\lambda = 1$ is the textual layer and $\lambda = 0$ is the pictorial layer. $\alpha > 0$ and $\beta > 0$ are parameters used to adjust the effect of the two components. In this work, we set $\beta = 1$ to simplify this definition, since the structural differences are important to both textual and pictorial regions. α is used to adjust the effect of the luminance component when the textual layers are processed. As illustrated in Fig.5.4, human beings are not sensitive to the colour intensity change derived from some degree of quantization or contrast change, we calculate the difference between the textual layers from the reference and the distorted SCI to measure the degree of these intensity change. The difference is measured as follows:

$$d = (2 \cdot v_1 \cdot v_2) / (v_1^2 + v_2^2); \quad (5.11)$$

where $v_1 = \max(X_t) - \min(X_t)$ and $v_2 = \max(Y_t) - \min(Y_t)$. The value of α can be

determined by d and a threshold δ as follows:

$$\alpha = \begin{cases} 0 & \text{if } \lambda = 0 \\ d & \text{if } \lambda = 1 \text{ and } d > \delta \\ 1/d & \text{if } \lambda = 1 \text{ and } d \leq \delta \end{cases} \quad (5.12)$$

Based on the calculated quality maps of textual layer Q^1_map and pictorial layer Q^0_map , the quality scores of the textual and pictorial regions are computed as the mean values of the corresponding regions, respectively:

$$Q_t = \frac{Q^1_map \cdot T}{M} \quad (5.13)$$

$$Q_p = \frac{Q^0_map \cdot (1 - T)}{N} \quad (5.14)$$

where $M = \text{sum}(T)$ is the number of textual pixels; $N = \text{sum}(1 - T)$ is the number of pictorial pixels. Cooperating with the proposed weighting model in SPQA, the final quality score of the distorted image Y is computed:

$$Q_Y = W_t * Q_t + W_p * Q_p \quad (5.15)$$

5.4 Experimental Results

In this section, we first verify the reliability of the subjective scores, and then test the validity of the proposed weighting model based upon the subjective scores. Finally, we investigate the effectiveness of the proposed SPQA scheme to assess the quality of the SCIs in the SIQAD.

5.4.1 Reliability of DMOS

When processing the raw subjective scores, outliers are firstly detected and rejected according to the method [56]. A raw score for an image is considered to be an outlier if the difference between the score and the mean value is larger than the standard deviation. And for any session, all quality evaluations of a subject will be rejected if more than 16 of his evaluations are outliers. Then, we examine the consistency of all subjects' judgements of each image. According to [62], the consistency can be measured by the confidence interval derived from the number and standard deviation of scores for each image. Generally, with a probability of 95% confidence level, the distribution of the scores can be regarded as reliable. After outlier rejection, DMOS values of all images

are computed and their confidence intervals are obtained. In Fig.5.6, two examples of DMOS distributions with 95% confidence interval are shown, which demonstrate the agreement of subjects on the visual quality of images.

Figure 5.6: Distributions of DMOS values of two examples. The error bars indicate the confidence intervals of related scores.

Generally, the quality scales of the distorted SCIs in the database should exhibit good separation of perceptual quality and span the entire range of visual quality (from distortion imperceptible to severely annoying) [119]. Fig.5.7 shows the histogram of the DMOS values (0:100) of all distorted images in the database. It can be observed that the DMOS values of images range from low to high, and have a good spread at different levels.

Figure 5.7: Histogram of DMOS values of images in the SIQAD.

5.4.2 Analysis of Subjective Scores of Different Regions

In this section, we preliminarily test the validity of the proposed weighting method (introduced in 5.3.1). After the subjective test, we get three subjective scores for each test image: QE , QT and QP , corresponding to the quality of the entire, textual and pictorial regions, respectively.

Firstly, we analyze the correlation of these three quality scores (QE , QT and QP) in terms of *Pearson Linear Correlation Coefficient* (PLCC), *Root Mean Squared Error* (RMSE) and *Spearman rank-order correlation coefficient* (SROCC) [120]. As such, we can roughly know which part attracts more attention of observers. Meanwhile, correlations for each distortion type are also calculated to estimate human visual perception to different distortion types. The correlation results are reported in Table 5.1. From Table 5.1, we can observe that the textual part has higher overall correlation with the entire image than the pictorial part. However, for different distortion types, the results vary to some extent. For example, in the CC case, the contrast variation of pictorial regions affect human vision more compared to that of textual regions. The reason is that, observers prefer to give high scores to texts of high shape integrity and clarity, even though their colors change significantly. For pictorial regions, severe contrast change would result in uncomfortable viewing experience. Therefore, in this case, pictorial regions contribute more to the quality of the entire image. On the contrary, in the MB case, textual regions attract more attention. The integrity and clarity of texts are easier to be affected by motion blurring. For other distortions, the correlation

Table 5.1: Correlation analysis of the obtained quality scores for the entire images, textual and pictorial regions.

	QE and QT			QE and QP		
Distortions	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC
GN	0.9749	2.7974	0.9571	0.9777	2.7885	0.9393
GB	0.9835	2.3815	0.9571	0.9665	3.0782	0.9482
MB	0.9749	2.1825	0.9475	0.9380	3.1350	0.9032
CC	0.9217	3.8243	0.8446	0.9457	3.0667	0.8746
JPEG	0.9542	2.3801	0.9000	0.8967	3.1158	0.8596
JPEG2000	0.9144	3.1033	0.8625	0.9082	3.1530	0.8589
LSC	0.9187	2.6754	0.8196	0.9002	2.9464	0.8464
Overall	0.9338	4.8067	0.9148	0.8833	6.2471	0.8620

results also vary from case to case. Consequently, it is a challenging problem to build an unified formula to account for the correlation among the three scores.

As an initial attempt towards solving this problem, we propose a weighting model for estimating the quality of the entire image based on the quality of textual and pictorial regions, as described in Sec.5.3.1. The performance of the proposed model is measured by computing the correlation between the estimated and ground truth scores. Meanwhile, we compare the proposed model with a simple averaging combination of textual and pictorial scores. Table 5.2 reports the comparison results. It shows that the results of the proposed model are more consistent with human visual perception. Although there is still space to improve the performance, the proposed prediction model reflects the contributions of textual and pictorial regions with a high reliability. We will further demonstrate the effectiveness of the weighting model by applying it to objective metrics in the following section.

5.4.3 Validation of the Proposed SPQA

In this section, we use the images in the *SIQAD* to conduct the comparison experiments by using the proposed SPQA and other existing ones. The following 11 state-of-the-art NIQA metrics are adopted: PSNR, SSIM [68], MSSIM [69], IWSSIM [121], VIF [71], IFC [70], VSNR [72], MAD [74], FSIM [73], GSIM [75] and GMSD [76]. These metrics are implemented using the codes on their websites. We apply all the metrics to the

Table 5.2: Comparison of two combination methods

	Average combination			Proposed prediction model		
Distortions	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC
GN	0.9048	5.2572	0.8707	0.9847	2.2691	0.9607
GB	0.9032	5.4064	0.8654	0.9833	2.3888	0.9554
MB	0.9005	5.8983	0.8464	0.9798	1.9622	0.9464
CC	0.8577	6.0412	0.8168	0.9573	2.7777	0.8732
JPEG	0.8609	6.0150	0.8382	0.9458	2.4178	0.9018
JPEG2000	0.8373	6.6196	0.8329	0.9372	2.6148	0.8946
LSC	0.8120	6.9176	0.8136	0.9141	2.6838	0.8536
Overall	0.8674	5.9514	0.8433	0.9472	3.8577	0.9234

grayscale version of images, and compute the correlations between the predicted scores and DMOS values in terms of PLCC, RMSE and SROCC. Meanwhile, the correlations of specific distortions are calculated, to investigate the effectiveness of objective methods for different distortion types. We set $c_1 = 0.0026$, $c_2 = 0.0062$, and δ is experimentally set to 0.95 in the experiments.

We report the correlation results in Table 5.3, 5.4 and 5.5, where the first two with the best performance are marked with the **bold** font. It is shown that the proposed SPQA achieves the highest overall correlation with DMOS values, no matter of PLCC, SROCC or RMSE. Correlations between the SPQA scores and DMOS values for different distortion types are distinct from each other, as most of the other metrics. Particularly, there are much higher values for the first three distortions (i.e. GN, GB and MB) than others. The reason is that observers are sensitive to such kinds of distortions allocated in the entire image, and are able to distinguish the images with different distortion levels. For the remaining four types, especially for the CC case, the correlation results are not so high. The reason is that the contrast change only affects the intensity of texts, but not the integrity of texts about which subjects care more. By contrast, the NIQA metrics take the intensity variation into account, resulting in the inconsistency with DMOS values. The proposed SPQA has taken this situation into account, and thus the predicted quality of all the test images has higher consistency with human perception compared with other existing metrics.

In Fig. 5.8, we also provide the scatter plots of the predicted quality scores against

(a) PSNR

(b) SSIM

(d) MSSIM

(e) VIF

(f) IFC

(h) FSIM

(e) GSIM

(h) SPQA

Figure 5.8: Scatter of predicted quality scores by some metrics against the DMOS values on the SIQAD. The vertical axis in each figure is the DMOS values.

the DMOS values for some representative objective metrics (such as PSNR, SSIM, MSSIM, VIF, IFC, FSIM, GSIM, GMSD and SPQA) on the SIQAD. The seven kinds of distortions (GN, GB, MB, CC, JPEG, JPEG2000 and LSC) are separatively displayed with different markers. From Fig. 5.8, it can be observed that the predicted scores by the SPQA have the most centralized distribution than others. In most of other metrics, the distribution of predicted scores on all distortion types is somehow dispersive. For example, for PSNR and GSIM, the distribution of predicted scores on the CC distortion deviates much from the distribution on other kinds of distortions, degrading their overall performance.

In Fig. 5.9, a reference SCI (a) and its several distorted versions (b)-(f) are given for visual quality comparison. We can see that, from (b) to (f), the DMOS values of these images are increasing, indicating the descending of the visual quality. However, the three measures (PSNR, SSIM and VIF) do not have the same changing tendency, which means that they cannot achieve high consistency with the DMOS values at these cases. These three metrics generally capture the practical variations occurring in the distorted images, without considering the different perception of viewers to different regions in SCIs. For instance, in the subjective test, observers prefer to give high scores to images with clear and unbroken textual regions, even though their intensity values

Table 5.3: Correlation results (PLCC) of the DMOS values and the objective scores given by 12 metrics. The paired T-test is applied to the proposed SPQA against the 11 NIQA methods. The result ($H = 1, P < 0.05$) for each pair indicates that the SPQA is significantly better than the tested 11 NIQA methods.

Metrics	GN	GB	MB	CC	JPEG	J2K	LSC	Overall
PSNR	0.9748	0.9802	0.9631	0.8542	0.9403	0.9096	0.9169	0.6244
SSIM	0.9668	0.9780	0.9648	0.9284	0.9231	0.9063	0.9192	0.7977
MSSIM	0.9626	0.9755	0.9604	0.9276	0.9169	0.9103	0.9169	0.6508
IWSSIM	0.9626	0.9758	0.9545	0.9228	0.9186	0.9102	0.9165	0.6818
VIF	0.9682	0.9797	0.9664	0.8806	0.9245	0.9090	0.9275	0.8429
IFC	0.9727	0.9788	0.9676	0.9321	0.9301	0.9113	0.9292	0.6736
VSNR	0.9722	0.9802	0.9657	0.8710	0.9006	0.8852	0.9075	0.6217
MAD	0.9719	0.9802	0.9635	0.6678	0.9345	0.9112	0.9129	0.6555
FSIM	0.9476	0.9771	0.9039	0.8632	0.9208	0.9068	0.9002	0.6073
GSIM	0.9636	0.9757	0.9596	0.9203	0.9207	0.9095	0.9147	0.6161
GMSD	0.9640	0.9808	0.9660	0.9225	0.9194	0.9082	0.9164	0.7542
SPQA	0.9624	0.9791	0.9653	0.9313	0.9231	0.9088	0.9162	0.8761

Table 5.4: Correlation results (SROCC) of the DMOS values and the objective scores given by 12 metrics.

Metrics	GN	GB	MB	CC	JPEG	J2K	LSC	Overall
PSNR	0.9375	0.9411	0.9375	0.7589	0.8625	0.8696	0.8268	0.6020
SSIM	0.9393	0.9464	0.9393	0.7196	0.8554	0.8679	0.8250	0.7897
MSSIM	0.9411	0.9536	0.9018	0.7821	0.8482	0.8714	0.8196	0.6345
IWSSIM	0.9446	0.9536	0.9357	0.7875	0.8446	0.8679	0.7929	0.6689
VIF	0.9393	0.9429	0.9393	0.7607	0.8536	0.8661	0.8268	0.8183
IFC	0.9321	0.9411	0.9393	0.8304	0.8536	0.8661	0.8161	0.6347
VSNR	0.9321	0.9411	0.9393	0.7321	0.8321	0.8607	0.8036	0.5933
MAD	0.9321	0.9411	0.9393	0.6089	0.8607	0.8821	0.8125	0.6398
FSIM	0.9286	0.9357	0.8804	0.7071	0.8589	0.8339	0.8232	0.5669
GSIM	0.9446	0.9375	0.9268	0.7625	0.8536	0.8429	0.7982	0.5832
GMSD	0.9393	0.9411	0.9411	0.8071	0.8482	0.8679	0.8071	0.7243
SPQA	0.9393	0.9411	0.9411	0.7964	0.8536	0.8768	0.8232	0.8521

Table 5.5: Correlation results (RMSE) of the DMOS values and the objective scores given by 12 metrics.

Metrics	GN	GB	MB	CC	JPEG	J2K	LSC	Overall
PSNR	2.8622	2.5150	2.8209	5.4663	2.6938	3.1672	2.5319	10.6303
SSIM	3.2204	2.6001	2.7770	3.9241	2.8237	3.2362	2.5445	8.1220
MSSIM	3.4264	2.7949	2.8931	3.6132	2.8978	3.1532	2.5881	10.3000
IWSSIM	3.4224	2.7566	3.0011	3.6757	2.8300	3.1813	2.6033	9.9270
VIF	3.1413	2.5691	2.7163	4.8653	2.8235	3.1791	2.4075	7.2295
IFC	2.9685	2.6505	2.6476	3.7642	2.7733	3.1239	2.3897	9.9235
VSNR	2.9364	2.5417	2.7122	4.6796	3.0444	3.3721	2.8048	10.6568
MAD	2.9883	2.5524	2.8222	7.6812	2.6562	3.1878	2.6078	10.2413
FSIM	3.7938	2.6528	3.5739	4.8120	2.8437	3.2697	2.6829	10.4559
GSIM	3.3648	2.7525	2.9539	3.8398	2.8771	3.1761	2.6014	10.6430
GMSD	3.3401	2.4724	2.7289	3.5697	2.8636	3.2101	2.5996	8.7898
SPQA	3.3593	2.6074	2.7600	3.4708	2.8304	3.1952	2.5637	6.4646

Figure 5.9: Visual quality comparison of SCIs with different distortion types. The DMOS values and the quality scores predicted by four different metrics (PSNR, SSIM, VIF and SPQA) are provided for comparison.

have been changed. Compared with images in (c)(d), the image (b) gets the highest visual quality. However, PSNR and SSIM values of (c)(d) are higher than those of (b). Additionally, most subjects have a bad impression to the blurring effect at the first sight, and thus give low scores to the blurred images. As shown in Fig. 5.9 (d)-(f), the images in (d) and (e) have better visual quality than image (f) with severe motion blur. However SSIM value of (e) and VIF value of (d) are lower. This phenomenon can also be observed in Fig.5.6, where most of the DMOS values for blurred images (from the first eight to the twenty-one points) are higher than other images.

5.4.4 Proposed Weighting Strategy on NIQA Methods

In Sec.5.4.2, we have demonstrated that the proposed weighting model performs well on estimating the overall quality score based on the scores of textual and pictorial regions. In Sec.5.4.3, the performance of the proposed SPQA including the weighting model has also been proved promising. In this section, we apply the proposed weighting model to some representative NIQA metrics, such as SSIM, VIF, FSIM and GMSD to verify the advantages of the proposed object metric. In particular, we substitute the proposed metric with such NIQA metrics in the SPQA scheme, evaluating the quality of two layers separately. The correlation between DMOS and predicted scores by the modified NIQA metrics (marked as S_{-}) is computed and reported in Table 5.6. From this table, we can see that the performance of some modified NIQA metrics are improved when the proposed weighting method is integrated, such as S_{SSIM} , S_{FSIM} and S_{GMSD} . But the improvement is still far away from satisfaction in evaluating the visual quality of distorted SCIs. As to the S_{IFC} and S_{VIF} , the performance drops somehow. In addition, we test the performance of the parameter setting of α in the proposed SPQA. We combine the luminance and sharpness similarity simply by $\alpha = \beta = 1$, and the experimental result is demonstrated as $SPQA_s$ in Table 5.6. From Table 5.6, we can see that, without the adaptive setting of α , the performance of $SPQA_s$ is not as good as that of SPQA. Overall, the proposed SPQA with the weighting strategy works much better than other relevant existing objective metrics.

Table 5.6: Comparison of two combination methods

	SSIM vs. S_SSIM		IFC vs. S_IFC		VIF vs. S_VIF	
PLCC	0.7977	0.8374	0.6736	0.5641	0.8429	0.8136
SROCC	0.7897	0.8283	0.6347	0.5544	0.8183	0.7927
RMSE	8.122	7.3725	9.9235	10.9946	7.2295	7.7885
	FSIM vs. S_FSIM		GMSD vs. S_GMSD		SPQA_s vs. SPQA	
PLCC	0.6073	0.6317	0.7542	0.7942	0.8473	0.8760
SROCC	0.5669	0.5982	0.7243	0.7692	0.8290	0.8520
RMSE	10.4559	10.3928	8.7898	8.1062	7.0254	6.4681

5.5 Summary

In this chapter, we have carried out an in-depth study on perceptual quality assessment of distorted SCIs, from both subjective and objective perspectives. A new large-scale image database, *SIQAD*, is built to explore the subjective quality evaluation of SCIs. DMOS values of images in the database are obtained via the 11-category ACR subjective testing, and their reliability is verified. Based upon the three subjective scores for textual, pictorial and overall regions, we find that textual regions contribute more to the quality of the entire image in most distortion cases. The proposed weighting model works well to account for this relationship. Combining with the weighting model, a new objective quality metric is constructed to separately assess the visual quality of textual and pictorial regions. The proposed integration scheme, named SPQA, outperforms existing 11 NIQA objective metrics on visual quality evaluation of distorted SCIs, which has been demonstrated by the experimental results.

Chapter 6

Conclusion

SCIs have different characteristics from natural scene images and scanned document images, and this poses as a challenge in designing algorithms for processing SCIs. In this thesis, we have exploited the basic properties of SCIs, and proposed the algorithms to address three important problems in processing SCIs, namely SCI segmentation, compression and visual quality assessment.

6.1 Research Contributions

In Chapter 3, we have thoroughly studied the text segmentation problem for SCIs. We have figured out the problems in the existing segmentation methods, based on which, we have proposed a coarse-to-fine segmentation framework. The coarse stage, i.e., the LIAM based enhancement, removes most of pictorial regions, only leaving few high frequency pictorial parts on the textual layer. We have also presented SOIG as another important component in the coarse-to-fine framework, which employs adaptive grouping technique to connect texts with various scales and orientations. The MAD and some morphological operations have been adopted to ensure that the generated candidate TCCs have uniform geometrical features. The proposed three verification criteria can filter out false positive TCCs and form a final clean textual layer. We have demonstrated that, the proposed text segmentation framework can maintain text integrity as well as avoid over- or under- segmentation of texts.

Another topic we concern in this thesis is SCI compression problem. In Chapter 4, we have proposed a learning based method for specially encode textual regions in

SCIs. A tailored text dictionary have been learned as the specialized basis functions for textual image coding, which can guarantee the sparse representation of SCIs. We have evaluated the sparsity and energy compaction of several over-completed basis with different number of items. The over-completed DCT based training procedure achieved faster convergence than original K-SVD that uses random samples as the initial dictionary. We have demonstrated that the proposed SCI coding scheme achieves higher compression performance than the standard coding methods, such as JPEG, JPEG2000 and DjVu, especially when encoding SCIs with large percentage of texts.

In Chapter 5, we have first carried out the investigation on perceptual quality assessment of SCIs. As the first step, we built a large scale image database SIQAD including 1000 SCIs. Serving as the ground truth to compare the effective of various objective metrics, subjective quality scores of images in the SIQAD have been obtained via our extensive subjective testing. The reliability of the obtained subjective scores has been verified. Specifically, three subjective scores were provided to entire, textual and pictorial regions of each SCI, respectively, based on which, we have revealed the relationship between text quality scores and entire image quality scores. According to this discovery, a novel weighting model has been proposed to account for the contribution of textual or pictorial regions to the final quality of entire SCI. We have verified that existing NIQA or DIQA methods are not applicable to assess visual quality of distorted SCIs. A specific objective metric, SPQA, has been proposed for automatically evaluating visual quality of SCIs. The SPQA takes into account the different visual response to textual and pictorial regions, thereby having high consistency with HVS as viewing SCIs.

6.2 Future Directions

As the first step, we investigated the objective quality measure in the full reference manner, exploring the similarity between reference and distorted SCIs. Currently, we designed the SPQA scheme based on the different response to textual and pictorial regions, by only considering the effect of luminance and sharpness change. In the future, we aim to build more effective metrics for assessing textual regions, by taking the sizes, percentages and representation modes of textual regions into account. The full reference quality assessment can help us to easily figure out which kind of information

is helpful in SCI quality assessment. Furthermore, based on knowledge of construction of full-reference quality metric, we will transfer it into a no reference manner to meet the requirement of more real-world applications.

Besides, quality assessment of SCIs can be further studied and applied to QoE of SCIs. The effect of some factors (e.g., text readability, text clarity or viewing comfort) to QoE should be clearly exploited. In the current work, we just give one score to summarize these factors. In the next step, we will investigate more about the effect of these factors separately, taking more text characteristics (e.g., typefaces, fonts, positions, area-ratios) into account. The QoE of SCIs can be further guided by these factors, and then applied to QoE related applications (such as Web QoE).

Our current study focuses on screen content images, just considering the textual characteristics in one image. In many of SCI related applications, such as remote education, virtualized screen sharing and computer screen monitoring, SCIs are transmitted in a sequence, which can be regarded as screen content videos. More benefits can be achieved if we consider more inter-frame properties of texts in sequenced SCIs. Hence, we will extend our current work to precess screen content video in the future. Firstly, we will propose novel visual quality assessment algorithms for screen content videos. Moreover, we will investigate how to adjust the coding procedure by using a perceptual video quality metric. We aim to propose an effective perceptual compression model to encode screen content videos.

Bibliography

- [1] Y. Lu, S. Li, and H. Shen. Virtualized screen: a third element for cloud-mobile convergence. In *IEEE Multimedia*, 2011. xi, 1, 3, 51, 63
- [2] T. Chang and Y. Li. Deep Shot: a framework for migrating tasks across devices using mobile phone cameras. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011. xi, 1, 3, 4, 51, 63
- [3] C. Yi and Y. Tian. Text string detection from natural scenes by structure-based partition and grouping. *IEEE Trans. on Image Processing*, 20(9):2594–2605, 2011. xii, 14, 28, 49, 51, 61
- [4] E. Haneda and C. Bouman. Text segmentation for MRC document compression. *IEEE Trans. on Image Processing*, 20(6):1611–1626, 2011. xii, 14, 15, 16, 28, 52, 55
- [5] H. Shen, Y. Lu, F. Wu, and S. Li. A high-performanance remote computing platform. In *IEEE International COnference on Pervasive Computing and Communications*, 2009. 1, 63
- [6] T. Lin and S. Wang. Cloudlet-screen computing: a multi-core-based, cloud-computing-oriented, traditional-computing-compatible paralled computing paradigm for the masses. In *IEEE Conference on Multimedia and Expo*, 2009. 1, 51
- [7] H.Shen, Y. Lu, F. Wu, and S. Li. Low cost realtime screen sharing to multiple clients. In *IEEE Conference on Multimedia and Expo*, 2010. 1

- [8] T. Lin and P. Hao. Compound image compression for real-time computer screen image transmission. *IEEE Trans. on Image Processing*, 14(8):993–1005, 2005. 1, 4, 11, 16, 22
- [9] C. Lan, G. Shi, and F. Wu. Compress compound images in H.264/MPGE-4 AVC by exploiting spatial correlation. *IEEE Trans. on Image Processing*, 19:946–957, 2010. 1, 4, 18
- [10] W. Ding, D.Liu, Y.He, and F. Wu. Block-based fast compression for compound images. In *IEEE Conference on Multimedia and Expo*, 2006. 1, 4, 16, 18, 28, 63
- [11] Z. Pan, H. Shen, S. Li, and N. Yu. A low-complexity screen compression scheme for interactive screen sharing. *IEEE Trans. on Circuits and Systems for Video Technology*, 23(6):949–960, 2013. 1, 4, 56, 82
- [12] Z. Pan, H. Shen, and Y. Lu. Brower-friendly hybrid codec for compound image compression. In *IEEE International Symposium on Circuits and Systems*, 2011. 1, 4, 16, 18, 28, 51, 63
- [13] S. Juliet and D. Florinabel. Efficient block prediction-based coding of computer screen images with precise block classification. *IET Image Processing*, 5(4):306–314, 2011. 1, 4, 16, 28, 51, 52
- [14] T. Richardson, Q. Stafford-Fraser, K. R. Wood, and A. Hopper. Virtual Network Computaing. *IEEE Internet computing*, 2(1):33–38, 1998. 2
- [15] Microsoft. Remote desktop protocol. [http://msdn.microsoft.com/en-us/library/aa383015\(v=vs.85\).aspx](http://msdn.microsoft.com/en-us/library/aa383015(v=vs.85).aspx). 2
- [16] H. Yang, W. Lin, and C. Deng. Learning based screen image compression. In *IEEE International Workshop on Multimedia Signal Processing*, 2012. 4
- [17] H. Yang, S. Wu, C. Deng, and W. Lin. Scale and orientation invariant text segmentation for born-digital compound images. *IEEE Trans. on Cybernetics*, 2014 to appear. 4
- [18] ISO/IEC JTC 1/SC 29/WG 11 Requirements subgroup. Requirements for an extension of HEVC for coding of screen content. In *MPEG 109 meeting*, 2014. 4, 5, 19

- [19] S. Hu, R.A. Cohen, A. Vetro, and C.-C.j. Kuo. Screen content coding for HEVC using edge modes. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013. 5, 19
- [20] G. Braeckman, S.M. Satti, H. Chen, S. Delputte, P. Schelkens, and A. Munteanu. Lossy-to-lossless screen content coding using an HEVC base-layer. In *18th International Conference on Digital Signal Processing (DSP)*, 2013. 5, 19
- [21] T. Lin, K. Zhou, X. Chen, and S. Wang. Arbitrary shape matching for screen content coding. In *Picture Coding Symposium (PCS)*, 2013. 5, 19
- [22] C. Chen, X. Xu, R. Liao, W. Peng, S. Liu, and S. Lei. Screen content coding using non-square intra block copy for HEVC. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2014. 5, 19
- [23] W. Zhu, W. Ding, J. Xu, Y. Shi, and B. Yin. Screen content coding based on HEVC framework. *IEEE Trans. on Multimedia*, 16(5):1316–1326, 2014. 5, 19
- [24] Z. Ma, W. Wang, M. Xu, and H. Yu. Advanced screen content coding using color table and index map. *IEEE Trans. on Image Processing*, 23(10):4399–4412, 2014. 5, 19
- [25] Z. Wang and A.C. Bovik. Mean Squared Error: Love It or Leave It? *IEEE Signal Processing Magazine*, 26:98–117, 2009. 5
- [26] W. Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: a survey. *Journal of Visual Communication and Image Representation*, 22:297–312, 2011. 5
- [27] T. Ciszkowski, W. Mazurczyk, and Z. Kotulski. Towards Quality of Experience-based reputation models for future web service provisioning. *Telecommunication Systems*, 51(4):283–295, 2012. 5
- [28] R. Schatz and S. Egger. An Annotated Dataset for Web Browsing QoE. In *6th International Workshop on Quality of Multimedia Experience*, 2014. 5
- [29] D. Guse, S. Egger, A. Raake, and S. Moller. Web-QoE under Real-world Distractions: Two Test Cases. In *6th International Workshop on Quality of Multimedia Experience*, 2014. 5

- [30] Damon M. Chandler. Seven Challenges in Image Quality Assessment: Past, Present, and Future Research. *ISRN Signal Processing*, 2013. 5, 20, 80
- [31] D. Karatzas, S. Mestre, J. Mas, F. Nourbakhsh, and P. Roy. ICDAR 2011 robust reading competition challenge 1: reading text in born-digital images (web and email). In *Conference on Document Analysis and Recognition (ICDAR)*, 2011. 11, 22, 45, 48
- [32] C. Jung, Q. Liu, and J. Kim. Accurate text location in images based on SVM output scores. *Image and Vision Computing*, 27:1295–1301, 2009. 14
- [33] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Conference on Computer Vision and Pattern Recognition*, 2010. 14, 28, 61
- [34] A. Gonzalez and L. Bergasa. A text reading algorithm for natural images. *Journal of Image and Vision Computing*, 31:255–274, 2013. 14, 28
- [35] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting texts of arbitrary orientations in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 14, 28
- [36] P. Shivakumara, T. Phan, and C. Tan. A laplacian approach to multi-oriented text detection in video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(2):412–419, 2011. 14, 28, 41
- [37] F.Y. Shih and S.S. Chen. Adaptive document block segmentation and classification. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 26(5):797–802, 1996. 14
- [38] M. Pietikainen and O. Okun. Edge-based method for text detection from complex document images. In *Sixth International Conference on Document Analysis and Recognition*, 2001. 14
- [39] Y. Zheng and D. Doermann. Machine printed text and handwriting identification in noisy document images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(3):337–353, 2004. 14

- [40] U.Pal and P. Roy. Multioriented and Curved Text Lines Extraction From Indian Documents. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(4):1676–1684, 2004. 14
- [41] Y. Chen and B. Wu. Text Extraction from Complex Document Images Using the Multi-plane Segmentation Technique. In *IEEE Conference on Systems, Man, and Cybernetics*, 2006. 14
- [42] C. Fung and R. Chamchong. A review of evaluation of optimal binarization technique for character segmentation in historical manuscripts. In *Conference on Knowledge Discovery and Data Mining*, 2010. 14
- [43] N. Francisco, N. Rodrigues, and E. Silva. Scanned compound document encoding using multiscale recurrent patterns. *IEEE Trans. on Image Processing*, 19:2712–2724, 2010. 14, 15, 16, 18, 51
- [44] A. Kapoor, P. Pandey, and K. Biswas. Fuzzy rule based document image segmentation for component labeling. In *Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2011. 14, 15, 28
- [45] K. Konstantinides and D. Tretter. A JPEG Variable Quantization Method for Compound Documents. *IEEE Trans. on Image Processing*, 9(7):1282–1287, 2000. 16, 18, 28, 63
- [46] W. Ding, Y. Lu, and F. Wu. Enable efficient compound image compression in H.264/AVC intra coding. In *IEEE Conference on Image Processing*, 2007. 16, 18, 28, 63
- [47] R. Queiroz, R. Buckley, and M. Xu. Mixed raster content (MRC) model for compound image compression. In *Conference on Visual Communications and Image Processing*, 1999. 17
- [48] ITU-T Recommendation. T.44 Mixed Raster Content (MRC). In *International Telecommunication Union*, 1999. 17
- [49] R. de Queiroz. Compressing compound documents. In *In the Document and Image Compression Handbook*, 2005. 17

- [50] Itu recommendation t.82, 1994. <http://www.itu.int/rec/T-REC-T.82-199303-I/en>. 17
- [51] L. Bottou, P. Haffner, P. Howard, P. Simard, Y. Bengio, and Y. LeCun. High quality document image compression with DjVu. *Journal of Electronic Imaging*, 7(3):410–425, 1998. 17
- [52] DjVu Technology. In <https://www.caminova.net/en/products/>, latest retrieval by 2014. 17
- [53] S.H. Wang and T. Lin. United Coding for Compound Image Compression. In *International Congress on Image and Signal Processing*, 2010. 18
- [54] A.K. Moorthy and A.C. Bovik. Visual quality assessment algorithms: what does the future hold? *Multimedia Tools and Applications*, 51(2):675–696, 2011. 20
- [55] S. Winkler. Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):616–625, 2012. 20
- [56] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. on Image Processing*, 15(11):3440–3451, 2006. 20, 81, 93
- [57] P. Le Callet and F. Autrusseau. Subjective quality assessment irccyn/ivc database, 2005. <http://www.irccyn.ec-nantes.fr/ivcdb/>. 20
- [58] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008A database for evaluation of full-reference visual quality assessment metrics. *Adv. Modern Radioelectron*, 10:30–45, 2009. 20, 81
- [59] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010. 20
- [60] D. Jayaraman, A. Mittal, A.K. Moorthy, and A.C. Bovik. Objective quality assessment of multiply distorted images. In *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 2012. 20

- [61] M. Chen, L.K. Cormack, and A.C. Bovik. No-Reference Quality Assessment of Natural Stereopairs. *IEEE Trans. on Image Processing*, 22(9):3379–3391, 2013. 20, 21
- [62] ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures. In *Int. Telecommunications Union*, 2012. 20, 80, 83, 85, 86, 93
- [63] H. Quan, M-N. Garcia, F. Speranza, P. Corriveau, and A. Raake. Study of rating scales for subjective quality assessment of high-definition video. *IEEE Trans. on Broadcasting*, 57(1):1–14, 2011. 20
- [64] R.K. Mantiuk, A. Tomaszewska, and R. Mantiuk. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491, 2012. 20
- [65] J. Li, M. Barkowsky, and P. Le Callet. Subjective assessment methodology for preference of experience in 3DTV. In *IEEE Image, Video and Multidimensional Signal Processing Technical Committee Workshop on "3D Image/Video Technologies and Applications"*, 2013. 20
- [66] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao. Hodgerank on random graphs for subjective video quality assessment. *IEEE Trans. on Multimedia*, 14(3):844–857, 2012. 20
- [67] I. Sprow, D. Kuepper, Z. Baranczuk, and P. Zolliker. Image quality assessment using a high dynamic range display. In *International AIC Color Congress*, 2012. 21
- [68] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004. 21, 89, 96
- [69] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *Conference record of the thirty-seventh asilomar conference on signals, systems and computers*, 2003. 21, 96

- [70] H.R. Sheikh, A.C. Bovik, and G.de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. on Image Processing*, 14(12):2117–2128, 2005. 21, 96
- [71] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Trans. on Image Processing*, 15(2):430–444, 2006. 21, 96
- [72] D.M. Chandler and S.S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. on Image Processing*, 16(9):2284–2298, 2007. 21, 96
- [73] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: a feature similarity index for image quality assessment. *IEEE Trans. on Image Processing*, 20(8):2378–2386, 2011. 21, 96
- [74] E.C. Larson and D.M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electron Imaging*, 19(1):011006:1–011006:21, 2010. 21, 96
- [75] A. Liu, W. Lin, and M. Narwaria. Image quality assessment based on gradient similarity. *IEEE Trans. on Image Processing*, 21(4):1500–1512, 2012. 21, 89, 96
- [76] W. Xue, L. Zhang, X. Mou, and A.C. Bovik. Gradient magnitude similarity deviation: an highly efficient perceptual image quality index. *IEEE Trans. on Image Processing*, 23(2):684–695, 2014. 21, 89, 96
- [77] Q. Li and Z. Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):202–211, 2009. 22
- [78] M. Lin, S. Li, F. Zhang, and K. Ngan. Reduced-reference image quality assessment using reorganized DCT-based image representation. *IEEE Trans. on Multimedia*, 13(4):824–829, 2011. 22
- [79] R. Soundararajan and A.C. Bovik. RRED indices: reduced reference entropic differencing for image quality assessment. *IEEE Trans. on Image Processing*, 21(2):517–526, 2012. 22

- [80] J. Wu, W. Lin, G. Shi, and A. Liu. Reduced-reference image quality assessment with visual information fidelity. *IEEE Trans. on Multimedia*, 15(7):1700–1705, 2013. 22
- [81] M.A. Saad, A.C. Bovik, and C. Charrier. Blind image quality assessment: a natural scene statistics approach in the DCT domain. *IEEE Trans. on Image Processing*, 21(8):3339–3352, 2012. 22
- [82] L. He, D. Tao, X. Li, and X. Gao. Sparse representation for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 22
- [83] W. Xue, L. Zhang, and X. MOu. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 22
- [84] P. Ye, J. Kumar, and D. Doermann. Beyond human opinion scores: blind image quality assessment based on synthetic scores. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 22
- [85] W. Xue, X. Mou, L. Zhang, A.C. Bovik, and X. Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Trans. on Image Processing*, 23(11):4850–4862, 2014. 22
- [86] P. Ye and D. Doermann. Document image quality assessment: a brief survey. In *International Conference on Document Analysis and Recognition*, 2013. 22
- [87] I. Guyon, R. M. Haralick, J. J. Hull, and I. T. Phillips. Data sets for OCR and document image understanding research. In *Proceedings of the SPIE - Document Recognition IV*, 1997. 22
- [88] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *ACM SIGIR conference on Research and Development in Information Retrieval*, 2006. 22
- [89] H. Hase. Quality evaluation of character image database and its applications. In *International Conference on Document Analysis and Recognition*, pages 1414–1418, 2011. 22

- [90] J. Kumar, P. Ye, and D. Doermann. Diqa: Document image quality assesment datasets, 2013. <http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=73>. 22
- [91] P. Ye and D. Doermann. Learning features for predicting OCR accuracy. In *International Conference on Pattern Recognition*, pages 3204–3207, 2012. 22
- [92] D. Kumar and A. G. Ramakrishnan. QUAD: Quality assessment of documents. In *International Workshop on Camera-based Document Analysis and Recognition*, pages 79–84, 2011. 22
- [93] T. Obafemi-Ajayi and Gady Agam. Character-based automated human perception quality assessment in document images. *IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(3):584–595, 2012. 22
- [94] A. Mittal, R. Soundararajan, and A.C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letter*, 20(3):1–4, 2013. 23, 24
- [95] H. Yang, W. Lin, and C. Deng. Image Acitivity Measure (IAM) for Screen Image Segmentation. In *IEEE International Conference on Image Processing*, 2012. 24, 86, 88
- [96] J. Song, Z. Li, M.R. Lyu, and S. Cai. Recognition of Merged Characters Based on Forepart Prediction, Necessity-Sufficiency Matching, and Character-Adaptive Masking. *IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics*, 35(1):2–11, 2005. 28
- [97] T. Singh, S. Roy, O. Singh, T. Sinam, and Kh. Singh. A new local adaptive thresholding technique in binarization. *International Journal of Computer Science Issues*, 8:271–277, 2011. 32, 45
- [98] ‘Closing’. In <http://homepages.inf.ed.ac.uk/rbf/HIPR2/close.htm>, latest retrieval by 2014. 35
- [99] H. Freeman and R. Shapira. Determining the minimum-area encasing rectangle for an arbitrary closed curve. *Communications of the ACM*, 18(7):409–413, 1975. 37

- [100] Skeletonization. In <http://cs.smith.edu/~nhowe/research/code/>, latest retrieval by 2014. 41
- [101] BCID database. In <https://sites.google.com/site/yanghuanhomepage/>. 45, 51
- [102] S. Shah and R. Vemuri. An analysis on the effect of image features on lossy coding performance. *IEEE Signal Processing*, 7(5):104–107, 2000. 45
- [103] L. Li and Z. Wang. Compression quality prediction model for JPEG2000. *IEEE Trans. on Image Processing*, 19:384–398, 2010. 45
- [104] Textdetector. <http://www.cvc.uab.es/icdar2011competition/?com=results>. 48
- [105] R. de Queiroz. On data-filling algorithms for MRC layers. In *IEEE International Conference on Image Processing*, 2000. 56
- [106] K. Jung, K. Kim, and A.K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, 37:977–997, 2004. 60
- [107] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for Sparse Representation Modeling. *Proceedings of IEEE*, 98(6):1045–1057, 2010. 64
- [108] J. Mairal, F. Bach, and J. Ponce. Online dictionary learning for sparse coding. In *Proceedings of the ICML*, 2009. 64
- [109] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher Discrimination Dictionary Learning for Sparse Representation. In *International Conference on Computer Vision*, 2011. 64
- [110] J. Wright, Y. Ma, J. Mairal, T. Huang, and S.C. Yan. Sparse Representation for Computer Vision and Pattern Recognition. *Proceedings of IEEE*, 98(6):1031–1044, 2010. 65
- [111] M. Aharon, M. Elad, and A. M. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. on Signal Processing*, 54(11):4311–4322, 2006. 66, 68
- [112] M. Aharon. Overcomplete Dictionaries for Sparse Representation of Signals. In *PhD thesis, The Technion Israel Institute of Technology*, 2006. 66

- [113] G. Reeves, V. Ullal, and E. Battenberg. Sparse Signal Representation: Image Compression using Sparse Bayesian Learning. 2006. 68
- [114] Image Compression with Set Partitioning in Hierarchical Trees. In <http://www.cipr.rpi.edu/research/SPIHT/>. 71
- [115] SIQAD. <https://sites.google.com/site/subjectiveqa/>. 81
- [116] L. Li and Z.S. Wang. Compression Quality Prediction Model for JPEG2000. *IEEE Trans. on Image Processing*, 19(2):384–398, 2010. 88
- [117] Y.H. Lee, J.F. Yang, and J.F. Huang. Perceptual activity measures computed from blocks in the transform domain. *Signal Processing*, 82:693–707, 2002. 88
- [118] R. Hassen, Z. Wang, and M. Salama. Image sharpness assessment based on local phase coherence. *IEEE Trans. on Image Processing*, 22(7):2798–2810, 2013. 91
- [119] K. Soundararajan, NR. Soundararajan, A. C. Bovik, and L. K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Trans. on Image Processing*, 19(6):14271441, 2010. 94
- [120] Final report from the video quality experts group on the validation of objective models of video quality assessment. <http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx>. 95
- [121] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Trans. on Image Processing*, 20(5):1185–1198, 2011. 96

Publications

JOURNAL ARTICLES

1. **Huan Yang**, Shiqian Wu, Chenwei Deng, and Weisi Lin. Scale and Orientation Invariant Text Segmentation for Born-Digital Compound Images. *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp.533-547, Mar 2015.
2. **Huan Yang**, Yuming Fang, Yuan Yuan, and Weisi Lin. Subjective Quality Evaluation for Compressed Digital Compound Images. *Journal of Visual Communication and Image Representation*, Vol. 26, pp:105-114, Jan 2015.
3. **Huan Yang**, Yuming Fang, Weisi Lin. Perceptual Quality Assessment of Screen Content Images, *IEEE Transactions on Image Processing*, 2015(to appear).

CONFERENCE PAPERS

1. **Huan Yang**, Yuming Fang, Weisi Lin, and Zhou Wang. Subjective Quality Assessment of Screen Content Images. In *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX'14)*, Singapore, September, 2014.
2. **Huan Yang**, Weisi Lin, Chenwei Deng, and Long Xu. Study on Subjective Quality Assessment of Compressed Digital Compound Images. In *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS'14)*, Melbourne, Australia, June 2014.
3. **Huan Yang**, Weisi Lin, and Chenwei Deng. Image Activity Measure (IAM) for Screen Image Segmentation. In *Proceedings of the 19th IEEE International Conference on Image Processing (ICIP'12)*, Orlando, FL, USA, September 2012.

4. **Huan Yang**, Weisi Lin, and Chenwei Deng. Learning Based Screen Image Compression. In proceedings of the 14th IEEE International Workshop on Multimedia Signal Processing (MMSP'12), Banff, Canada, September 2012.